

**DEVELOPMENT OF AN EXPERT JUDGEMENT ELICITATION
AND CALIBRATION METHODOLOGY FOR RISK ANALYSIS IN
CONCEPTUAL VEHICLE DESIGN**

ODU PROJECT NO: 130012
AGENCY NO: NASA NCC-1-02044

FINAL REPORT:

**DEVELOPMENT OF AN EXPERT JUDGEMENT ELICITATION
METHODOLOGY USING CALIBRATION AND AGGREGATION FOR
RISK ANALYSIS IN CONCEPTUAL VEHICLE DESIGN**

JANUARY 2004

Submitted by:

Resit Unal, Charles Keating
Bruce Conway and Trina Chytka

Engineering Management Department
Old Dominion University, Norfolk, VA 23529

Submitted to:

Roger A. Lepsch

Vehicle Analysis Branch

Aerospace Systems, Concepts and Analysis Competency
NASA Langley Research Center, Mail Stop 365
Hampton, Virginia 23681

EXECUTIVE SUMMARY

A comprehensive expert-judgment elicitation methodology to quantify input parameter uncertainty and analysis tool uncertainty in a conceptual launch vehicle design analysis has been developed. The ten-phase methodology seeks to obtain expert judgment opinion for quantifying uncertainties as a probability distribution so that multidisciplinary risk analysis studies can be performed. The calibration and aggregation techniques presented as part of the methodology are aimed at improving individual expert estimates, and provide an approach to aggregate multiple expert judgments into a single probability distribution. The purpose of this report is to document the methodology development and its validation through application to a reference aerospace vehicle. A detailed summary of the application exercise, including calibration and aggregation results is presented. A discussion of possible future steps in this research area is given.

TABLE OF CONTENTS

EXECUTIVE SUMMARY.....	II
INTRODUCTION.....	1
DEVELOPMENT OF EXPERT ELICITATION QUESTIONNAIRE.....	4
<i>POPULATION</i>	<i>6</i>
<i>CALIBRATION OF ASSESSMENTS</i>	<i>8</i>
<i>CALIBRATION ALGORITHM DEVELOPMENT</i>	<i>9</i>
<i>EXPERT CALIBRATION FUNCTION VALIDATION</i>	<i>10</i>
<i>CALIBRATION FUNCTION RELIABILITY</i>	<i>11</i>
AGGREGATION PROCESS.....	14
APPLICATION OF EXPERT JUDGMENT ELICITATION METHODOLOGY.....	18
<i>WEIGHTS AND SIZING CASE DESCRIPTION</i>	<i>19</i>
<i>OPERATIONS SUPPORT CASE DESCRIPTION</i>	<i>19</i>
<i>QUESTIONNAIRE DESIGN, IMPLEMENTATION AND DEPLOYMENT.....</i>	<i>20</i>
RESULTS.....	29
<i>Calibration</i>	<i>29</i>
<i>Aggregation.....</i>	<i>31</i>
VALIDATION.....	33
DISCUSSION	38
CONCLUSION.....	42
EXTENSIONS OF RESEARCH	43
REFERENCES	58
 APPENDICES	
Appendix A: Expert Judgment Sample Questionnaire	45
Appendix B: Calibrated Distributions.....	46
Appendix C: Aggregation Results.....	48
Appendix D: BESTFIT® Distribution by Variable	55
Appendix E: Graphical Representation of Select Aggregated Responses	56

LIST OF TABLES

Table 1: Characteristics of an Expert	7
Table 2: Weighting Factors for Weights & Sizing.....	25
Table 3: Weighting Factors for Operations Support	26
Table 4: Operations Support VAR01 Calibrated distribution.....	26
Table 5: Calibrated distributions with Weighting Factors	27
Table 6: Operations & Support VAR01 Summary Data.....	32

LIST OF FIGURES

Figure 1: Expert Judgment Elicitation Methodology Summary	3
Figure 2: Expert Elicitation Questionnaire Instructions.....	5
Figure 3: Calibration Technique	9
Figure 4: Instructions for Parameter Impact Classification.....	20
Figure 5: Background Section of Questionnaire (Weight & Sizing)	24
Figure 6: Background Section of Questionnaire (concluded).....	25
Figure 7: Operations Support VAR01 assessments	27
Figure 8: Operations Support VAR01 Excel View.....	28
Figure 9: Operations Support VAR01 Aggregated results.....	31
Figure 10: Validation Triad.....	35

LIST OF EQUATIONS

Equation 1: Calibration Factors.....	10
Equation 2: Linear Opinion Pool Algorithm.....	12
Equation 3: Linear Opinion Pool Equation	27
Equation 4: @RISK® Simulation Equation.....	28

INTRODUCTION

The National Aeronautics and Space Administration (NASA) Langley Research Center has long been responsible for advanced aerospace vehicle conceptual development. In determining attributes for an advanced concept vehicle, NASA utilizes various resources. Among these are current designs and technology, extrapolated to address future requirements and anticipated technology levels. The process of extrapolating current technology requires engineering judgment, and the degree to which the projections will be borne out is dependent upon the expertise of those forecasters performing the extrapolations.

In addition to ascertaining what technologies can or should be included in an advanced vehicle concept, it is important that the cost impact (positive or negative) of incorporating unproven technology be determined. The technologies under consideration can cover many disciplines and affect most, if not all, of the proposed vehicle's systems and subsystems (see, for example, Rowell, Olds, and Unal, 1999). In fact, adoption of a specific technology may impact more than one subsystem, to differing degrees. Because conceptual design specialists do not usually have expertise in every single vehicle system and their technologies, they may not always be able to judge accurately the projected impacts of incorporating future technology. Thus, a methodology to systematically guide the technology forecasting and assess the cost impact of adopting advanced concepts would be very desirable.

Expert judgment can provide useful information for forecasting, assessing risk and making decisions. Application areas in expert judgment have been diverse, including nuclear engineering, various types of forecasting (economic, meteorological, technical,

etc), military intelligence, seismic risk and environmental risk from toxic chemicals.

With the rise in use of expert opinion to assess risk in high consequence environments, methods are needed in order to combine judgments when multiple experts are queried for their opinions. The use of multiple experts is often employed when very little “hard data” is available and/or when multiple disciplines are involved in the analysis. The prime concern of most researchers in risk analysis when using multiple experts is how the multiple opinions should be combined or aggregated to ensure adequate capture of diverse judgments (Morgan and Henrion, 1990).

Prior to aggregating multiple opinions, assessments need to be elicited from the experts. To reduce the variability (uncertainty) of the assessment values, a calibration function is applied. Calibration can help in reducing the effects of base-rate fallacy and overconfidence (Ayyub, 2001). Many authors have discussed calibration of experts in judgment elicitation scenarios. Morgan and Henrion (1990) note that there have been many empirical studies focused on the calibration aspects of people’s abilities as probability assessors. Keren (1991) points out that most of the calibration studies he reviewed focused on technical formal issues, presumably because the dominant perspective is that uncertainty is a reflection of the external world and thus of the events or outcomes being assessed.

Through the review of literature on uncertainty, expert judgment elicitation, calibration, and aggregation, a methodology has been developed that utilizes characteristics of an expert elicitee to adjust the rendered judgments. These adjusted, or calibrated judgments lead to uncertainty distributions that appear to provide a more consistent response among multiple experts. The calibrated judgment distributions are

subsequently used as inputs to the aggregation process. The output of the aggregation process provides decision makers with a succinct representation of design uncertainty.

Figure 1 presents a summary of the expert judgment elicitation methodology, from problem definition to validation. Succeeding chapters of this report will describe various elements of the methodology and the results of applying the methodology to a pilot case.

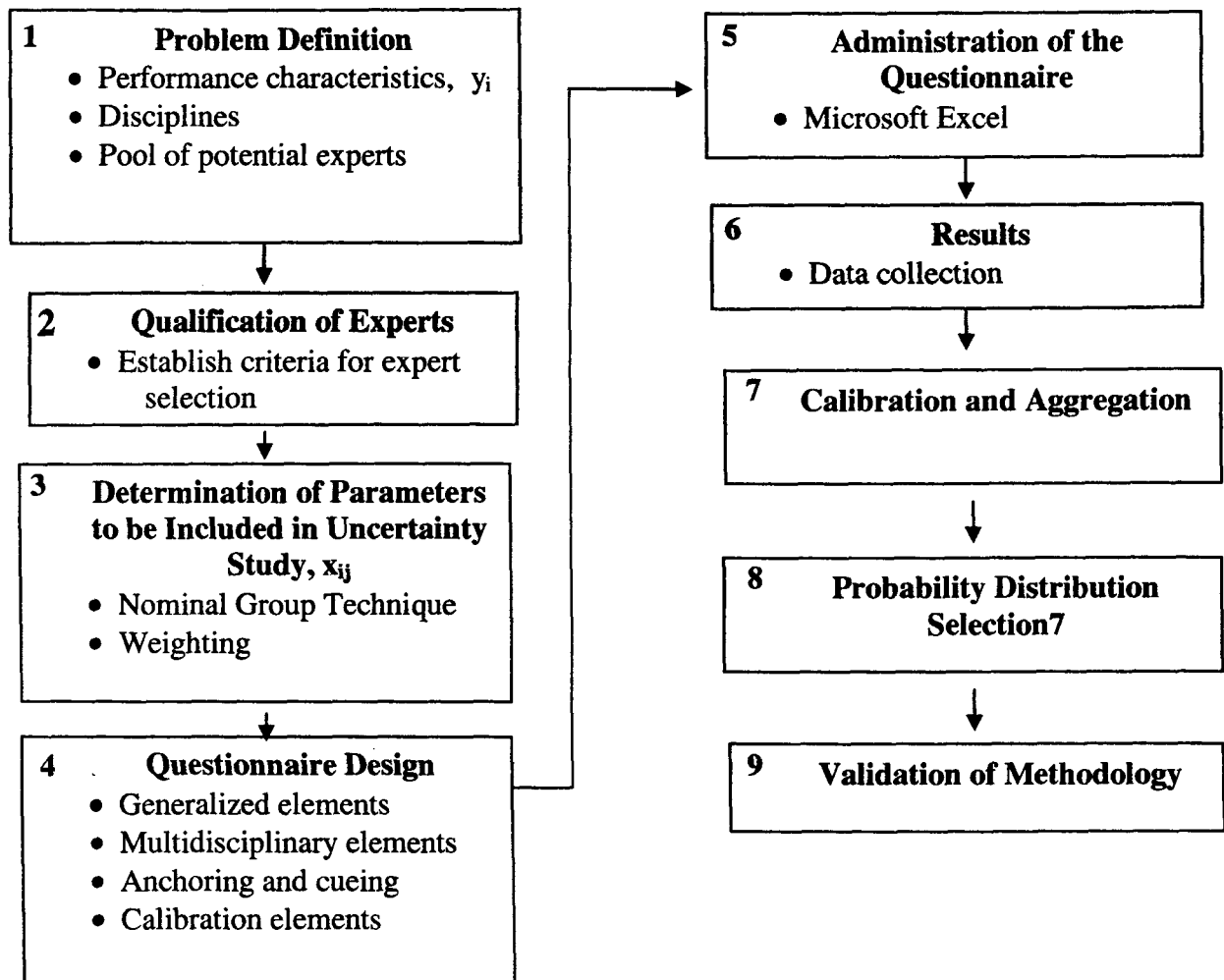


Figure 1: Expert Judgment Elicitation Methodology Summary

DEVELOPMENT OF EXPERT ELICITATION QUESTIONNAIRE

The questionnaire methodology developed for this research application expands upon the work of Monroe (1997). The questioning style of Monroe is that of an open-end structure allowing the expert to answer questions freely and provide anchoring and cueing descriptions. Monroe also adopted Likert scaling for his questionnaire using the 5-point scaling system for both quantitative and qualitative ratings. The Monroe methodology also assumes a default symmetrical triangular distribution associated with an expert's assessed uncertainty about a parameter of interest. For the current research, the Monroe questionnaire methodology was adopted and modified in two distinct ways. First, the format for which respondents were able to provide the high and low values for a variable of interest was more explicitly detailed. The Monroe questionnaire, as structured, allowed the respondents to contradict their evaluations of uncertainty around a variable of interest. Peterson (2000) asserts that a properly structured questionnaire prohibits contradiction and ambiguity. The questionnaire structure was altered to reflect clarity in defining the uncertainty of the variable. Figure 2 presents the set of instructions for the experts responding to the questionnaire; in particular, instruction 3 reflects the primary modification made to Monroe's approach.

A list of (discipline specific model) input parameters whose values are potentially uncertain will be provided on a subsequent screen. You will be asked to evaluate these parameters using the following guidelines.

1. Rate each INPUT parameter uncertainty QUALITATIVELY using a 5-point rating scale (Low, Low/Moderate, Moderate, Moderate/High, High). Focus only on those INPUT parameters that you feel should be evaluated in this manner.
2. If you feel a parameter's default value should be modified, you may provide a new point estimate for the nominal value.

3. If you feel the range of possible values (due to uncertainty, physical limitations, design constraints, etc.) around the nominal value is not symmetrical, please provide your own estimates of minimum and maximum values.
4. Describe the reason for the uncertainty and the reasoning behind the parameter value ranges for the UNCERTAIN INPUTS that you rated. Include a rationale for those parameters to which you have assigned new nominal values. Do this simultaneously while rating each INPUT parameter to document your thinking.
5. Think of any other cues (or reasons that you have not documented) and record that information at this time.
6. Once the INPUT parameters provided have been rated for uncertainty, you may add parameters not shown which you assess to have a level of uncertainty associated with their value. Use the OTHER option listed at the bottom of the INPUT parameter listing for this purpose.
7. After rating all INPUT parameters, next anchor your Low, Moderate, and High QUALITATIVE measures of uncertainty to QUANTITATIVE measures on the 5-point scales (provided).
8. Describe any scenarios that may change INPUT parameter values. Provide the alternate INPUT parameter values that in your judgment would be appropriate for the scenario

Figure 2: Expert Elicitation Questionnaire Instructions

Secondly, a Background section was added to the questionnaire to allow for the calibration of expert assessments prior to the aggregation process. The Background section was added in support of the research efforts of Conway (2003) who developed an Expert Calibration Function to reduce the variability (uncertainty) of the assessment values. The input distributions for the aggregation algorithm are the calibrated distributions resulting from the Expert Calibration Function applied to the raw distributions from the questionnaire responses. Development of the Background section and calibration rationale is described later in this report.

The context of the questionnaire centers on the elicited assessments from various disciplinary experts (E_{ij}) on design parameters evaluated most prone to uncertainty. The experts are queried to evaluate input design parameters (x_{ij}) as well as uncertainty associated with the analysis tools (z_{ij}) they employ in the conceptual design process. The questionnaire participants are asked to provide their assessments of the level of uncertainty associated with the parameter (high, medium or low) and to adjust the nominal value provided if they feel the value is inaccurate. The nominal value (or

adjusted value if provided) coupled with the level of uncertainty assigned to the parameter will determine the probability distribution assigned to the design parameter $[(x_{ij}), (z_{ij})]$ from expert (E_{ij}) . The questionnaire results from each disciplinary expert are the input distributions to the Expert Calibration Function developed by Conway (2003). The output of applying the calibration function to the distributions results in reduced variability distributions, which, in turn, are the inputs into the aggregation process.

The platform from which the questionnaire is launched is a Microsoft Excel® (Microsoft Corporation, 2000, Version 9.0 3821 SR-1) workbook. The Excel workbook includes a “tab” spreadsheet for instructions, a “tab” for a sample questionnaire, a “tab” for the Background section, multiple “tabs” for the variables of interest, and a “tab” for tool uncertainty assessment. The use of multiple “tabs” within the Excel shell enables the questionnaire to be exported to the experts in one compact file which makes for ease of use and practicality. The questionnaire is electronically mailed (e-mailed) to each Subject Matter Expert (SME) for the respective disciplines. The advantage of using e-mail to distribute the questionnaire is that it allows the experts to assess uncertainty ratings on their own time and in a familiar setting – their workspace.

Population

The target population for this questionnaire methodology is the pool of NASA aerospace engineers and design manager teams in multiple NASA locations. The chosen participants are recognized experts in their respective fields of study. In the present instance, familiarity with multidisciplinary launch vehicle design and optimization are also a key criteria. The selection of appropriate subject matter experts by the design managers will be guided by the adherence to characteristics of expertise assembled from the literature. Much of the literature on identification of expertise (Shanteau, et al., 2002;

Jackson, 1999; Ayyub, 2001) asserts that no one criteria should be used as a selection basis or disqualifier for the identification of an expert. For example, the literature does not support that “x” number of years of experience or “y” minimum educational background is used explicitly as selection criteria for the identification of experts. While there has been some positive correlation between years of experience or educational background, there is no evidence to support applying this standard universally (Conway, 2003). The number of years of experience, educational background, cognitive skills, etc. are criteria to be integrated together in the selection process. No one criterion is considered a disqualifier for expertise; expertise is an integrated summation of the characteristics (criteria) described. The design team managers were given instructions to reflect upon the selection criteria listed in Table 1 as an integrated compilation of characteristics of an expert and then identify discipline specific experts based upon their subjective assessments of an individual’s expertise.

Expert Characteristics
Domain knowledge ▪ Years of experience ▪ Educational background
Cognitive skills ▪ Ability to discern usefulness of data
Decision strategies
Expert-task congruence ▪ Appropriate expertise for discipline specific task

Table 1: Characteristics of an Expert

Calibration of Assessments

The thrust of this part of the research effort was the development of calibration algorithms to apply to elicited expert judgment information on operation and support, weight and size, and multidisciplinary system requirements for advanced launch vehicles. The purpose is to aid designers in determining a “best” expert estimate of the value of discipline-related parameters in achieving projected performance in the realization of new systems. For example, in the operations and support discipline, improved supportability implies a reduction in the number of failures recorded against a system (measured as a percentage) that then require maintenance actions in order to return the system to flight readiness. It also implies the same reduction in the time and manpower required to maintain and service the system.

Many expert judgment elicitation scenarios involve events whose occurrence can be validated, because they are either past events or near term future events. In such cases, calibration of the expert assessors can include feedback on their performance, which could be expected to improve future performance (self-calibration). In the present research problem application, however, the preponderance of occurrences being assessed are in the distant future – as much as 20 or 30 years. Feedback involving actual results or occurrences is impossible. It is imperative, then, that a calibration technique be found for use by decision makers that does not rely on feedback for credible estimates. Thus, in the present research, an external calibration based on a pre-elicitation calibration questionnaire was sought. Figure 3 presents a schematic of this concept.

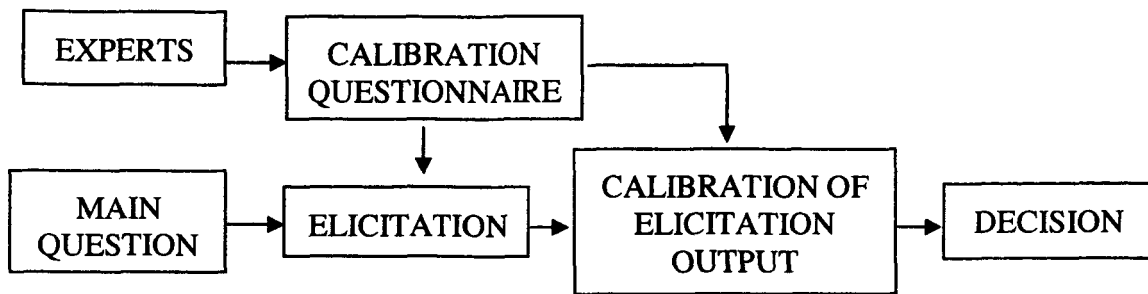


Figure 3: Calibration Technique

Calibration Algorithm Development

Fuzzy logic and Bayesian statistical techniques were employed to develop an Expert Calibration Function (ECF) based on degree (level and time) of past experience and current philosophy. For this study, the ECF has been developed for experts in several technology areas associated with advanced launch vehicles. A simple questionnaire was designed to “pigeon-hole” a responding expert into one of a set of experience classification categories, essentially a self-designation of expertise (Wright, Rowe, Bolger and Gammack, 1994). A second part of the questionnaire attempts to place the expert in his or her natural confidence level category, such as overconfident (presumes higher success probability than is actually achieved), underconfident (presumes lower), or neutral (places the correct probability). This was achieved through the use of utility theory and the outlining of several “wagers” (or choices of options) related to topics with which the experts are familiar. The questionnaire (and the validation discussed in the next section) were administered to a pilot group of several experts at the Langley Research Center.

From the experience and philosophy responses, calibration factors are determined such that an adjusted probability distribution for the expert’s uncertainty for each

parameter or analysis tool considered in the elicitation questionnaire can be subsequently constructed. The adjustment takes the following general form, where E is expertise, P is philosophy (confidence level), μ and σ^2 are statistics from the parameter uncertainty distribution, and A1 and A2 are arbitrary constants (which were set to 1 for this study) of the adjustment relations:

$$\Delta\mu = f(E, P, \mu)A_1$$

$$\Delta\sigma^2 = f(P, \sigma^2)A_2$$

Equation 1: Calibration Factors

It should be noted that the adjustment factors so determined will only be placeholder estimates until validated (and possibly modified) through a validation procedure as outlined next.

Expert Calibration Function Validation

Validation of the calibration methodology and resulting calibration functions is accomplished through an interrater comparison between initial and calibrated results from multiple experts participating in trial testing of an overall expert judgment elicitation, calibration, and aggregation methodology. One of the principal motivators for the current research was the (sometimes wide) disparity among multiple experts addressing the same uncertainty-related questions. A successful reduction in disparity among expert respondents' results would suggest at least partial validation of the calibration methodology. Further support would be provided by "movement" of the results from respondents with a somewhat lower level of expertise toward results from respondents possessing a higher level of expertise.

Calibration Function Reliability

Consistency of the expert calibration function's performance, or reliability, is expected to be high, given the mathematical nature of its form. This assumes that interpretation by experts of questions in the Background section of the expert elicitation instrument is consistent. Care was taken in the phrasing of the questions, and formulation followed guidance from the literature on similar constructions (see, for example, Monroe, 1997 and Duarte, 2001). Administration of the questionnaires was such that each participating expert responded without consultation or influence (bias) from other experts.

AGGREGATION OF MULTIPLE EXPERT ASSESSMENTS

The deployment of an aggregation methodology to combine multiple expert opinions in a conceptual design environment is not straightforward nor has there been experimentation in this context to test one technique's compatibility with conceptual environments over another. Most combination techniques rely on an expert's prior distribution of the variable of interest or on known distributions for the variable to calculate likelihood functions and determine an experts' credibility. Neither of these two elements are available in the current research domain. Additionally, with no clear consensus among the literature of an aggregation methodology explicitly applicable to the attributes embedded in the conceptual design environment, the selection of an appropriate combination method is indeed the preference of the analyst. The attributes of the current research case coupled with extensive literature review supports the use of a mathematical aggregation technique, specifically a technique from the opinion pool sector. The linear opinion pool (also known as weighted average) is the most simplistic of the mathematical approaches and merely a weighted linear combination of each expert's probability assessment:

$$p(\theta) = \sum_{i=1}^n w_i p_i(\theta)$$

Equation 2: Linear Opinion Pool Algorithm

where n is the number of experts, θ is the unknown variable of interest, $p_i(\theta)$ represents expert i 's probability distribution and , $p(\theta)$ represents the combined probability distribution and the weights w_i sum to one. The weights (w_i) assigned to each probability represent the relative quality of assessment assigned to each expert. The

linear opinion pool method can be generalized to provide a broad set of combination rules, however, it does not allow for convenient representation of dependence among experts' judgments (Genest and Zidek, 1986). Preliminary assessment reveals the linear opinion pool presents the most straightforward method of combining the opinions of the experts.

A common argument to the use of the linear opinion pool is the selection of the weighting values, which are assigned to the experts' probability distribution. The weighting assignments may be based purely on the subjectivity of the decision maker and his/her assessment of the reliability of the expert in estimating values or it may be based upon proven correlation between past performance of predictability. Correlation between past performances of predictability is typically not available for experts working in conceptual environments since most conceptual experts work in the domain of feasibility not in prototype and production (Morris, 2003). For the aerospace industry's conceptual design sector, for which this research case is embedded, experts rarely, if ever, see their designs materialize from paper to prototype therefore assessing predictability confidence cannot be empirically determined. The weighting assignments for this research case are therefore based purely on the subjectivity of the design manager (decision maker).

AGGREGATION PROCESS

The design of the aggregation process is rather intuitive. The inputs for aggregation are the calibrated design parameter uncertainty assessments provided by each discipline specific expert. The experts are queried for their assessments of parameter uncertainty via a questionnaire and the expert assessments are calibrated using a calibration function derived from the answers in the Background section of the questionnaire. The calibrated distributions are input into a Microsoft Excel spreadsheet to be read into a risk assessment model. The calibrated uncertainty assessments are then used as input distributions to the aggregation process.

Prior to the aggregation of the calibrated distributions, appropriate weighting factors need to be applied to each derived distribution. The weighting factor reflects the perceived credibility of the expert by the decision maker. Since empirical statistics are not available from which to derive expert credibility, design managers will be asked to provide credibility ratings for each expert. For the current research case, discipline specific design managers will be queried to provide credibility ratings (weighting factors) which will be applied to the distributions of the respective expert. The elicitation of weighting factors by the design managers will be via an interview process in which the design managers will be given a list of the experts who will provide uncertainty assessments for design variables. The design managers, in the presence of a facilitator, will rank the experts in terms of reliability of prediction on a scale from 0 to 1. For example, discipline *Design Manager A* may rate two experts who will provide uncertainty assessments for design variables for discipline A. *Design Manager A* may rank *Expert 1* with a 0.70 credibility rating while *Expert 2* may receive a 0.30 credibility

rating. The combined weighting factors for all experts being assessed by the design manager must equal 1.0. The design manager may base these weighting factors on subjective reasoning and his/her knowledge of each assessor's expertise, experience, and prediction capability (Morgan and Henrion, 1990).

The next task in the aggregation process is to import the calibrated uncertainty distributions and the experts weighting factors into an aggregation platform. The risk assessment model, @RISK[®] will be utilized within the Microsoft Excel shell to perform mathematical aggregation using the linear opinion pool algorithm. @RISK[®] allows for the specification of 37 distribution types including Beta, Erlang, Gamma, Normal, Triangular, Uniform, etc. For the current research case, uncertainty assessments are queried in triangular distribution form – the experts assess the minimum, most likely and maximum values for a parameter of interest. The calibration function applied to the initial assessments do not alter the distribution structure therefore the calibrated distributions are also in triangular form. The @RISK[®] function '*=RiskTriang(minimum, most likely ,maximum)*' will convert values in spreadsheet cells into a triangular distribution from which weighting factors and combination algorithms can be applied. The application of the aggregation algorithm will result in a “most likely” value of the combined assessments. In order to determine the minimum and maximum values of combined responses, sampling of the distributions must be performed.

The sampling simulation module within @RISK[®] will sample from each of the expert assessed uncertainty distributions and apply the appropriate weighting factor to each distribution during the sampling process. The advantage of using the @RISK[®] software is that either a Monte Carlo or Latin Hypercube Sampling technique can be

chosen for the sampling process thus providing a more robust aggregated response. Monte Carlo Sampling is the more traditional sampling technique and is entirely random. Monte Carlo algorithms are available for all the distributions considered feasible for expert assessment (Vose 2000) however, Monte Carlo suffers from efficiency issues. To increase the accuracy of Monte Carlo simulations Morgan and Henrion (1990) advocate increasing sample size. Consequently to improve the accuracy of a Monte Carlo simulation, a large number of iterations are typically required. This can become quite problematic when computing resources and/or time are limited. One of the variance reduction techniques employed to reduce the number of iterations required to improve computation efficiency is known as Latin Hypercube Sampling. The principle behind Latin Hypercube Sampling is stratification of the input probability distributions. Stratification divides the distribution into equal segments and a sample is then taken randomly from each segment. In this method, sampling is forced to represent values in each interval and thus, is forced to recreate the input probability distributions (Palisade 2002). Latin Hypercube Sampling provides for faster run times by requiring less iteration for convergence. For this reason, Latin Hypercube Sampling will be the choice sampling technique for this research case.

The @RISK® software also has an embedded module called BESTFIT® which allows a user to call up the BESTFIT® macro and have the software perform goodness-of-fit tests on the sampled results and fit the most appropriate probability distribution to the resultant data. Once the simulation has been performed using the Latin Hypercube Sampling technique, the distribution fitting solution BESTFIT® takes the sampled data and finds the distribution function that best fits that data. BESTFIT® tests up to 26

distribution types using advanced optimization algorithms. Results are displayed graphically and through a statistical report including goodness-of-fit statistics.

Once aggregation of multiple distributions has been applied and Latin Hypercube Sampling performed to determine a resultant distribution, the BESTFIT® module will be used in this research to determine the most compatible aggregated probability distribution for the sample data. The selected distribution will be based upon goodness-of-fit statistics; in particular a Chi-Squared statistic will be used as a goodness-to-fit measure. The Chi-Squared Statistic is the best-known goodness-to-fit statistics and can be used with both continuous and discrete sampled data. A weakness of the chi-squared statistic is there are no clear guidelines for selecting the number and location of the bins (Palisade 2002). To minimize this weakness, @RISK® has an option that allows the user to set the number of bins to “Auto” and set the bin style to “Equal Probabilities”. These choices will be selected for the simulation.

APPLICATION OF EXPERT JUDGMENT ELICITATION METHODOLOGY

Working with program management from NASA Langley Research Center, two aerospace vehicle disciplines and a concept design vehicle were selected for the application of the aggregation methodology. The development and deployment of an aggregation methodology was in conjunction with a larger research effort incorporating expert judgment elicitation and calibration research. In the larger context, an expert judgment elicitation methodology including background data on experts for the purpose of calibration has been developed. The requirements specified by the Institutional Review Board for the protection of experimental subjects were achieved through careful design and deployment of the questionnaire instrument. The current mechanism for distribution of the data acquisition instrument (questionnaire) is through electronic mail however; the questionnaire is capable of being administered via the World Wide Web.

The vehicle chosen for application of the aggregation methodology is a Two-Stage-To-Orbit (TSTO), staged at Mach 3, conceptual vehicle. This vehicle has a First Stage Booster and a Second Stage Orbiter, both stages having highly uncertain design variables. Two disciplines were chosen to elicit uncertainty assessments for input variables, Weights and Sizing and Operations Support. The Weights and Sizing discipline had 2 identified subject matter experts who were selected by the design managers based upon the expertise criteria outlined. The Operations Support discipline had 3 identified subject matter experts based upon the same expertise criteria.

Weights and Sizing Case Description

NASA Langley Research Center utilizes a Configuration Sizing program (CONSIZ) to size a vehicle and determine the weights of subsystem components. CONSIZ is a program developed specifically at Langley Research Center and provides capability of sizing and estimating weights for a vehicle based upon Weight Estimating Relationships (WERs) derived from historical regression using Shuttle data, finite element analysis and technology readiness level. The CONSIZ program has a predefined initial list of user-defined parameters which make up the input variables to the program. Additional design parameters are necessary to run a CONSIZ model but are provided as “pass through” variables from other discipline applications. For the TSTO staged at Mach 3 vehicle; the Booster has 104 input variables, 54 of which are user defined and the Orbiter has 109 input variables – 58 user defined.

Operations Support Case Description

For Operational Support analyses, NASA Langley Research Center utilizes a Reliability and Maintainability Analysis Tool (RMAT) to calculate and assess vehicle maintenance burden, ground processing times and manpower requirements for conceptual vehicles. The underlying algorithms for the RMAT computations are regression models built from historical aircraft maintenance data and extrapolations to meet technology readiness level. RMAT is a complex, stand-alone, operational analysis code requiring expert user inputs (Unal, 2002). RMAT utilizes over 200 user defined input variables for an analysis and like CONSIZ, RMAT performs analysis on each element of the vehicle configuration.

Questionnaire Design, Implementation and Deployment

Discipline design team managers agreed that to ensure efficiency in questionnaire content and to reduce the time burden for the experts to provide uncertainty assessments, only those design inputs having the most impact on vehicle performance or operational support need to be queried. To this end, a modified Nominal Group Technique (NGT) was utilized to elicit the most highly uncertain design parameters for each discipline. As discussed, the Nominal Group Technique requires that participants meet in the same location and rank order alternatives synchronously. The modified NGT developed for this research does not require experts to evaluate and rank alternatives synchronously but allows them to rank alternatives at their workstations. The tally of alternatives for the modified NGT is identical to the NGT; the only deviation from common procedure is the allowance of experts to rank alternatives in the privacy of their own workspace.

The modified NGT was implemented by listing all discipline specific user inputs (112 for Weights and Sizing, 200 for Operations Support) in an Excel spreadsheet. Each discipline design team member was given an electronic version of the Excel spreadsheet along with the instructions presented in Figure 4.

Instructions for Classification of Parameter Impact

1. Please examine the list of 'user input' variables provided.
2. On a scale from 1 to 5 (5 least significant – 1 most significant) please rate each of the variables according to your assessment of impact significance on performance characteristics.

Figure 4: Instructions for Parameter Impact Classification

Once the design team managers completed the rankings, the results were tallied.

Design team managers supported the use of the Pareto Principle as the discretionary

paradigm from which the tally list would be reduced therefore, adhering to the Pareto Principle, the 20 percent of responses deemed to have the most impact on vehicle design or operational support were selected for inclusion in the uncertainty expert elicitation questionnaire

From the reduced list of input parameters, discipline specific questionnaires were constructed. The questionnaires were electronically mailed to each subject matter expert with a request to complete the questionnaire and return it electronically to the researcher within five days of receipt. The five-day time limit was simply the discretion of the researchers but it did afford the experts sufficient time to complete the questionnaire and not feel too hurried. Each expert, working independently was asked to evaluate each design variable for uncertainty. The expert was first asked to rate the degree of uncertainty associated with the design parameter based on a qualitative 5 unit rating scale (Low, Low/moderate, Moderate, Moderate/high, High). Next, the expert was asked to evaluate the nominal value provided for the variable – the expert could accept this nominal value or provide a value he/she believed more appropriately represented the nominal value for the test case parameter. Additionally, the expert was given an opportunity to establish a non-symmetrical distribution around the nominal value if he/she felt it appropriate.

The expert was next asked to describe the reason for the uncertainty rating for the parameter and the resultant parameter ranges if they were modified. The expert was asked to provide rationale for those parameter values that were altered. Additionally, the expert was asked to provide any other cues or insights into his/her logic and record that information in the block provided.

Once all input parameters had been assessed, the expert was given the opportunity to add parameters not shown which he/she believed to have a level of uncertainty associated with their value. After rating all input parameters, the experts were asked to anchor their qualitative measure of uncertainty to a quantitative value using the 5-point scale provided.

Application of the Calibration Methodology

Working with program officials from NASA Langley Research Center, two aerospace vehicle disciplines and an example conceptual design case were selected to apply the calibration technique in conjunction with other ongoing expert judgment elicitation and aggregation research. In this larger scale research effort, a survey for eliciting expert judgment for the selected disciplines has been developed. Included in the questionnaire development is elicitation of background data on experts for the purpose of calibration. The satisfaction of Institutional Review Board requirements for protection of experimental subjects was achieved through careful design and handling of the questionnaire instruments. The final survey design is capable of being administered to selected experts via the World Wide Web.

The administration of the overall expert judgment elicitation questionnaire, including calibration-specific questions, was accomplished by querying discipline experts at NASA Langley Research Center, Vehicle Analysis Branch. The questions were administered using a Microsoft Excel® spreadsheet, on which responses were entered for subsequent data collection and analysis. Because expert participants will likely be in geographically dispersed locations, and responding to the expert elicitation questionnaire by the several experts involved will be asynchronous, use of web-based tools is deemed crucial to the efficient collection of information. Accordingly, automated web-based

survey software, Inquisite® (Catapult Systems, Austin TX), has been used to develop a web-based version of the expert judgment elicitation questionnaire for future application. A web based version was developed as a proof of concept but was not used in this study.

Calibration Questionnaire Design and Implementation

Specific questions were designed to be used in a Background section of the expert judgment elicitation questionnaire, based on previously noted findings from the literature review. To ascertain level of expertise, questions to ascertain the participating expert's self-assessment of his own expertise were posed, per Wright, Rowe, Bolger and Gammack (1994). Another background question asked the responding expert to compare his degree of expertise in the discipline being addressed with those of his peers in the discipline. This was intended to provide a second indicator of the expert's self-designated level of expertise related to a more absolute scale. Also, age was included as a requested background response, in accordance with findings by MacCrimmon and Wehring (1986) and Crawford and Stankov (1996) that expertise can be related to the age of an elicitee.

Several Background questions attempted to place an expert on a continuum with respect to his confidence level, or comfort with expert judgments rendered in response to an elicitation. The first of these was included as the second part of a question designed to gauge an expert's knowledge of the discipline by asking a discipline-specific question (with a numerical answer) that practitioner experts would be able to answer. Another set of questions was designed to help ascertain the expert's assessment of his attitude or philosophy with respect to manifesting confidence in judgments made in his specific field

(discipline). These questions' purpose was to develop a baseline to help gauge response to the final set of Background entries.

The last Background questions consisted of a series of options for which the participating expert was asked to make a choice. The available choices for each set of options were designed to reflect (1) a more risky situation whose choice would imply a tendency toward overconfidence, or (2) a less risky situation whose choice would signal a tendency away from overconfidence and toward neutrality or underconfidence.

The relatively brief Background section to the expert judgment elicitation questionnaire provided the necessary input to develop a calibration function for the responding expert. The complete Background section is shown in Figure 5.

<i>BACKGROUND (Weight and Sizing Specific)</i> 1. Name or USERID: _____ 2. Your age _____ 3. In this subject area, rate your own level of expertise on a scale of 1 (least) to 5 (most) _____ 4. Think of others with similar experience working in this discipline. On a scale of 1 (much less than peers), to 3 (about the same), to 5 (much more than peers), how would you compare yourself to your peers with respect to expertise? _____ 5. Payload mass fraction for the Space Shuttle is _____. Your assessment of the probability that your estimate is correct: _____. [Discipline-specific] 6. Think about predicting weights of hardware system elements; do you usually predict more than actually occurs (5), less than actually occurs (1), or about the amount/number of times that actually occur (3)? ____ 7. In estimating in your subject area in areas that have associated uncertainty, do you think it is better to be (a) close to the actual value without a lot of confidence in the estimate, or (b) not very close to the actual value, but with a high degree of confidence in your estimate? ____ 8. In making estimates related to weight and sizing model input parameters, would you say you were, (a) usually right-on with a high degree of confidence, (b) right-on without a high degree of confidence, or (c) not very close but with a high degree of confidence, or (d) not very close, and with not much confidence _____ For the following pairs of choices, please select the one in each pair that is most comfortable or appealing to you:

Figure 5: Background Section of Questionnaire (Weight & Sizing)

9.
 - (a) Setting, in advance, the completion date for a multi-year project
 OR
 - (b) Establishing, in advance, technical milestones for a multi-year project
10.
 - (a) Estimating, in advance, total cost outlays for a multi-year project
 OR
 - (b) Identifying, in advance, cost elements for a multi-year project
11.
 - (a) Identifying, at conceptual design review, utilization scenarios for the successful project
 OR
 - (b) Predicting, at conceptual design review, technical performance characteristics of the completed hardware

Figure 6: Background Section of Questionnaire (concluded)

Aggregation Process

The discipline specific design managers (often the elicited experts themselves) were then asked to provide credibility assessments for the discipline experts providing the uncertainty assessments. For Weights and Sizing and Operations Support the design managers and the elicited experts were synonymous. A facilitator interviewed each discipline specific design manager separately and asked for his/her credibility ranking for each of the experts elicited for responses. The results of the interview process follow:

	Design Manager A	Design Manager B
Expert A	0.30	0.70
Expert B	0.30	0.70

Table 2: Weighting Factors for Weights & Sizing

	Design Manager A	Design Manager B	Design Manager C
Expert A	0.40	0.40	0.20
Expert B	0.40	0.40	0.20
Expert C	0.40	0.40	0.20

Table 3: Weighting Factors for Operations Support

Calibrated distributions were imported into the @RISK® software in basic form (minimum, most likely, maximum values) and triangular distributions were built for each variable assessed for uncertainty using the “RiskTriang” function.

<u>Variable</u>	<u>Var Name</u>	<u>Nom Value</u>
VAR 01	Sched Hrs	114750

Expert	MIN	MOST LIKELY	MAX
Expert A	40,000.0000	120,487.5000	125,000.0000
Expert B	90,000.0000	135,460.3618	140,000.0000
Expert C	90,000.0000	100,000.0000	110,000.0000

Table 4: Operations Support VAR01 Calibrated distribution

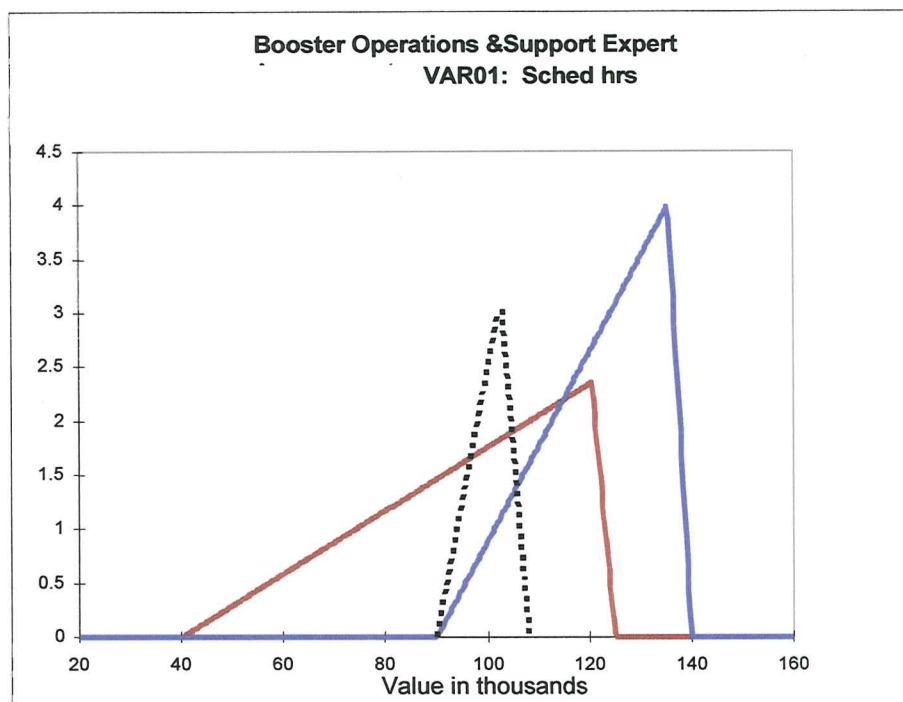


Figure 7: Operations Support VAR01 assessments

Weighting factors, as provided by the decision managers, were then applied to each of the expert's assessments.

	<u>Variable</u>	<u>Var Name</u>	<u>Nom Value</u>		
Weighting factor	VAR 01	Sched Hrs	114750		
	Expert	MIN	MOST LIKELY	MAX	
	0.40	Expert A	40,000.0000	120,487.5000	125,000.0000
	0.40	Expert B	90,000.0000	135,460.3618	140,000.0000
	0.20	Expert C	90,000.0000	100,000.0000	110,000.0000

Table 5: Calibrated distributions with Weighting Factors

A linear opinion pool aggregation algorithm was coded into a separate input cell.

Aggregated Weighted Distribution

$$= \text{RiskTriang}(C23,D23,E23)*A23 + \text{RiskTriang}(C24,D24,E24)*A24 + \text{RiskTriang}(C25,D25,E25)*A25$$

Equation 3: Linear Opinion Pool Equation

	A	B	C	D	E
20	Variable	Var Name	Nom Value		
21	VAR 01	Sched Hrs	114750		
22	Weighting factor	Expert	MIN	MOST LIKELY	MAX
23	0.4	Expert A	40000.0000	120487.5000	125000.0000
24	0.4	Expert B	90000.0000	135460.3618	140000.0000
25	0.2	Expert C	90000.0000	100000.0000	110000.0000

Figure 8: Operations Support VAR01 Excel View

The result of executing this equation is simply an aggregated “most likely” value of the combined distribution. For a more robust answer and to determine the minimum and maximum values of the distribution, a simulation approach is necessary. To perform simulation on the expert assessments, the addition of the variable “RiskOutput” to the aggregation equation is required and should not be considered part of the linear opinion pool equation itself.

Aggregated Weighted Distribution

=RiskOutput(“O&S Booster VAR01”)+ RiskTriang(C23,D23,E23)*A23+RiskTriang(C24,D24,E24)*A24+RiskTriang(C25,D25,E25)*A25

Equation 4: @RISK® Simulation Equation

To run a simulation, a variety of setting may be used to control the type of simulation @RISK® performs. A simulation in @RISK® supports unlimited iterations and multiple simulations (Palisade 2002). The “Simulation Settings” module allows you to specify the number of iterations to run as well as whether you want to use Monte Carlo or Latin Hypercube sampling. For this research case, 4 iterations were run (500, 1000, 10,000, 15,000) using the Latin Hypercube sampling technique. The number of iterations chosen was simply to monitor convergence of results and to compare aggregated values at different sampling iterations. Since increased iterations increases accuracy of simulated results (Vose, 2000), the results from running 15,000 iterations will be evaluated.

RESULTS

Calibration

Applying the calibration methodology to the experts' variable-related uncertainty responses was straightforward. The expert's initial uncertainty distribution was easily determined from his responses and displayed for subsequent comparison to the distribution obtained after applying a calibration adjustment, determined from responses to the background questionnaire for each expert.

The calibrated distributions follow the trends (compared to the initial distributions) suggested by the present research. For those experts whose confidence/risk philosophy tended toward risk-averse (denoted by negative philosophy scores, the calibrated distributions reflected a lower variance. There were no participating experts in either case whose philosophy score was positive, so observation of an expanded distribution was not possible. For the one expert with a zero philosophy score (in the operations and support case), the expected constancy of variance was observed in the case results.

The calibrated distributions also reflected the expected response to difference in expertise. Those experts with a higher expertise score displayed less adjustment in "most likely" values than did those with lower expertise scores. The shifts in modes also occurred in the direction established by the adjustment algorithm (which in turn is based on the expertise and confidence philosophy of each expert). There were responses that resulted in calibrated distributions that did not completely follow precisely those suggested by expertise and philosophy scores; these distributions were the result of an experts assignment of either an alternate "most likely" value to a parameter, or to the

assignment of minimum and/or maximum values that resulted in a nonsymmetrical distribution. As noted in an earlier discussion, endpoints so assigned to a distribution were honored in the subsequent analysis.

It is observed that applying the variance adjustments (based on the philosophy ratings of the experts) will tend to equalize variances among participants. This tendency reflects that associated with more traditional calibrations (that of measuring instruments, for example). A principal reason for applying calibrations is the reduction or elimination of measurement errors resulting from bias, thus rendering measurements more consistent from trial to trial. The use of multiple instruments to improve system redundancy often necessitates the aggregation of the output of these instruments to provide a “best” value for system operation. The removal of biases renders the aggregation task more easy, whether it be simple averaging or the use of more sophisticated weighting functions. In the present case, using the experts’ calibrated distributions for the subsequent aggregation process should likewise result in more consistent output for use in the disciplinary or multidisciplinary analysis upon which programmatic decisions may ultimately be based. Feedback of calibration results to participating experts is considered an important part of the methodology. Such feedback can perhaps indicate to an expert tendencies that could prove helpful in other analysis situations. The feedback will also allow the expert to contribute to the optimization and efficiency of the calibration methodology. Feedback and subsequent adjustment of the calibration techniques would be expected to occur over a period of time, since adjustments based on only a few respondents may not be representative those resulting from a larger pool of expert participants.

Aggregation

The numerical results of performing the aggregation process on the calibrated assessments are summarized in Appendix C. Minimum, most likely, maximum, and mode are represented. The resultant aggregated distributions are formulated by the module BESTFIT® which runs goodness-of-fit algorithms to determine the most compatible distribution that best represents the sampled data. As discussed earlier, the selected distribution from the simulated aggregation process will be based upon the Chi-Squared statistic. A graphical illustration with test statistics of an aggregated result follows:

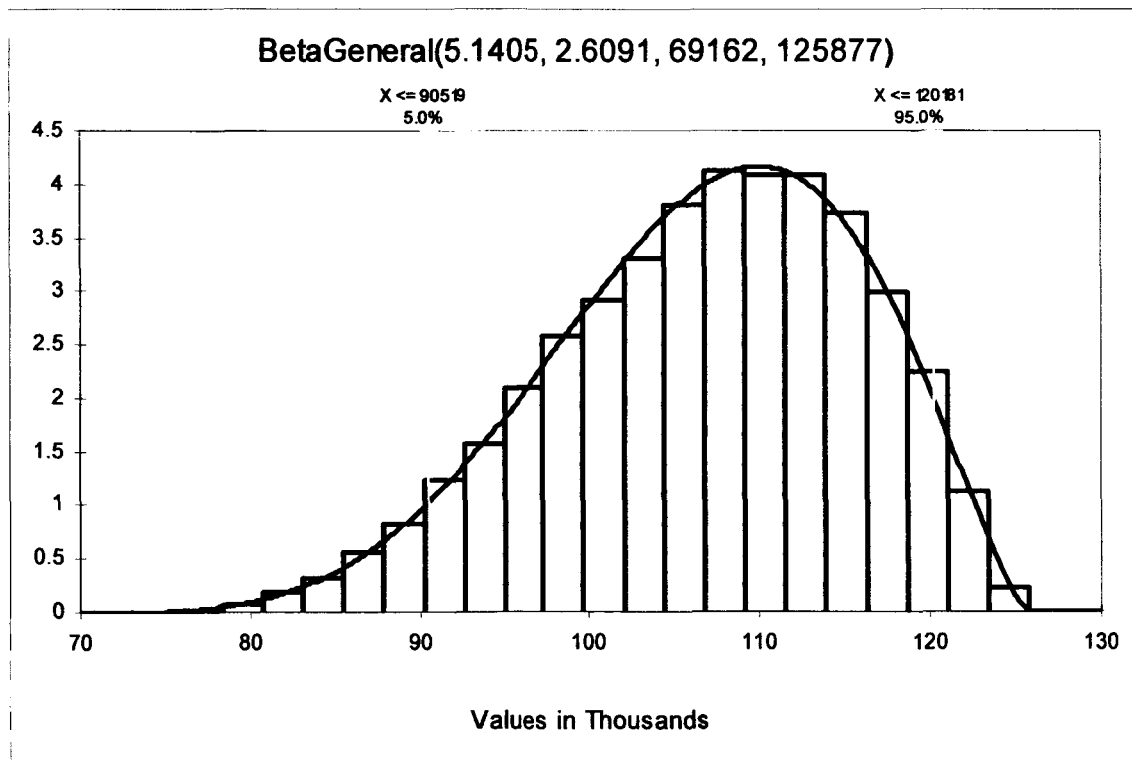


Figure 9: Operations Support VAR01 Aggregated results

Summary Data	
FIT	Beta Distribution
$\alpha 1$	5.153360407
$\alpha 2$	2.646831156
min	69354.10984
max	126007.711
Left X	90579
Left P	5.00%
Right X	120191
Right P	95.00%
Diff. X	2.96E+04
Diff. P	90.00%
Minimum	69354
Maximum	126008
Mean	106784
Mode	109922
Median	107595
Std. Deviation	9042.4
Variance	81765319
Skewness	-0.4109

Chi Square Statistics	
Test Value	107.5
P Value	0.0675
Rank	1
C.Val @ 0.75	77.7774
C.Val @ 0.5	86.3342
C.Val @ 0.25	95.4972
C.Val @ 0.15	100.6695
C.Val @ 0.1	104.275
C.Val @ 0.05	109.7733
C.Val @ 0.025	114.6929
C.Val @ 0.01	120.591
C.Val @ 0.005	124.7177
C.Val @ 0.001	133.5121
# Bins	88

Table 6: Operations & Support VAR01 Summary Data

Uncertainty assessments were queried for 33 variables (10 from Weights and Sizing and 23 from Operations Support) and the aggregation methodology was applied to the calibrated distributions. Certain variables were not aggregated because their values were identical in both the Orbiter and Booster assessments. Determining the most appropriate distribution to fit the aggregated responses resulted in 25 variables being most compatible with the Beta distribution, 2 variables best represented with the Triangular distribution and 1 aggregated variable best represented with a Weibull curve. A summary of the variables and the best-fit aggregated distribution are in Appendix D. Appendix E shows a select few graphical representations of the aggregated responses.

VALIDATION

Validation of engineering research has conventionally demanded “formal, rigorous, and quantitative validation” (Barlas & Carpenter, 1990). Traditional validation methods are based primarily on logical inductive and/or deductive reasoning which works well in predictable, stable, data rich environments. These validation methods have traditionally been classified as objective – based on deviance measures and statistical tests (Law & Kelton, 1991; Sargent, 1999). Objective methods align quite well with the empiricisms that “formal, rigorous and quantitative” validation demands. There are, however, areas of engineering research that rely on subjective assessments, which makes strict adherence to “formal, rigorous and quantitative” validation problematic. Science progresses, according to Thomas Kuhn (1970), when the ruling paradigms cannot provide adequate explanations to scientific problems under investigation and this inadequacy makes way for new paradigms. The inadequacy of objective methods in validating non-empirical environments gave rise to a new validation paradigm – the subjective method. Subjective validation, often used in knowledge based systems and simulation modeling (Sargent, 1999; Bawcom, 1997; Pederson et al., 2000) utilizes conversational, contextual and subjective validation. When observed data does not exist this method must obligatorily be used (Braga, unknown).

The validation method for the current research case is an adaptation of a method from the relativist validation paradigm referred to as the Validation Square– a method developed, deployed and validated by Pederson et al. The Validation Square is a validation method designed to evaluate the effectiveness and efficiency of a research method based on qualitative and quantitative measures The validation method set forth in

this research incorporates the principles of the Validation Square and adds another tier to the validation process – content validity of the elicitation instrument. The data acquisition instrument is a vital component to this research therefore the validation process would not be complete without a component to address instrument validity.

The validation of the methodology for risk assessment using expert elicitation is therefore, tri-fold. First, performance validity of the methodology is evaluated at the three levels defined in the Validation Square. An expert panel consisting of the design managers in an unstructured interview process (performed individually) qualitatively comment on (1) the acceptability that the outcome of the method is useful with respect to the initial purpose; (2) acceptability that the achieved usefulness is linked to applying the method; and (3) the usefulness of the method is beyond the case study. Second, the methodology itself must be proven to be compatible with conceptual vehicle design environments. Assessing the structural validity of the method is not empirically possible. There does not exist a validation data set from which to compare the results to an observed data point. The objective is to determine the usability and compatibility of the methodology based upon the confidence of design managers in its output. Therefore the method is determined defensible when it is shown the calibration and aggregation result is reproducible, accountable, subject to peer review, and unbiased. Structural validation is performed in an identical manner as performance validation – individual, unstructured, conversational interviews. Lastly, content validity of the data collection instrument is necessary. Evaluation of face validity through expert assessment is intended to determine that the questionnaire is precise, reproducible, and accountable. Figure 10 represents the validation triad.

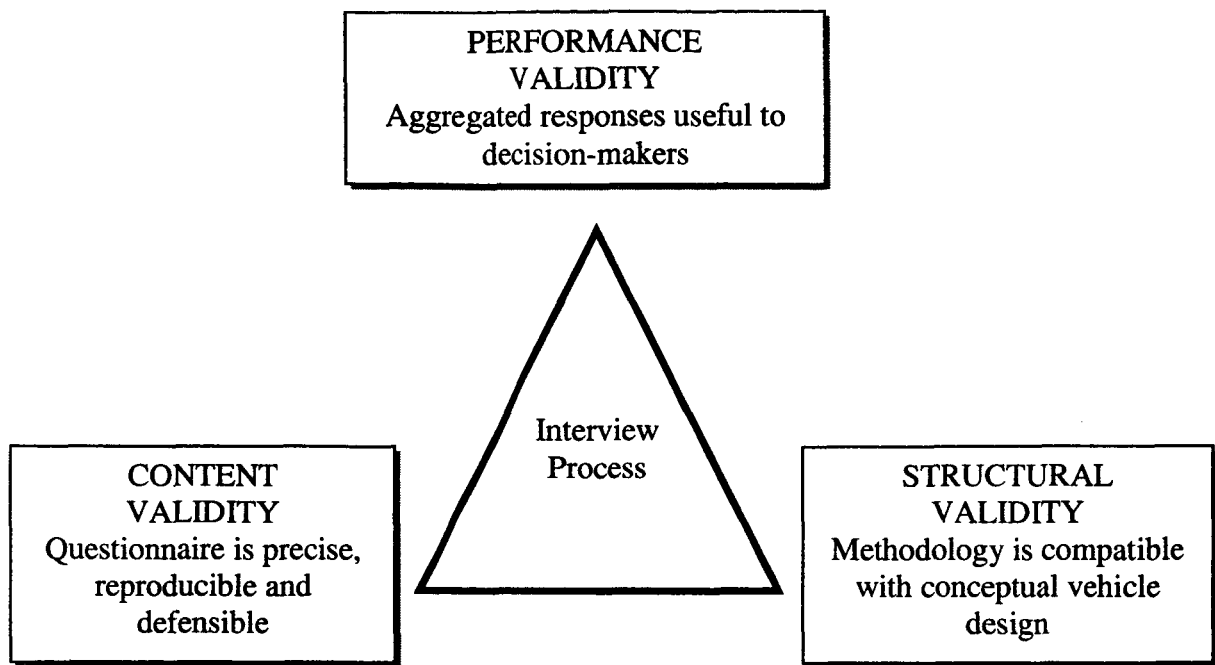


Figure 10: Validation Triad

Validation of the methodology is a one-on-one unstructured interview process consisting of a three-part construct. Content validity is assessed by the decision-makers in regards to ease of use of the questionnaire instrument, appropriateness of the questionnaire structure to the problem domain and comprehension of content and context of the questionnaire. Structural validity is assessed in regards to usability and value added of an aggregated response to decision strategies and the applicability of the method beyond the test case. Lastly, performance validity is based upon feedback from the decision-makers on which uncertainty representation they find most useful in their decision strategies.

The discipline design managers were separately interviewed and allowed to discuss any aspect of the methodology. A researcher guided the discussion when necessary to ensure the minimum requirements were covered. The researcher recorded

all responses and comments. The design managers unanimously agreed the questionnaire content, structure and deployment was efficient, well developed and user friendly. Each concurred the questionnaire could easily be applied to other test cases and successfully captured the qualitative and quantitative uncertainty of design parameters. Two specific suggestions for improving the questionnaire were reported by the design managers. The design managers suggest that instead of having the subject matter experts tab through the variables on Microsoft Excel[®] worksheet tabs, automate the process so that when the experts finish assessing the uncertainty of a variable, the sheet immediately scrolls to the next variable. Additionally, the design managers would like to see the questionnaire transported into the World Wide Web environment. Both of these suggestions have been previously identified in the larger context research from which this methodology is embedded and are recommended to be implemented as an extension of this research.

Determining whether the aggregated responses are useful to decision-makers (structural validity) and which uncertainty representation decision-makers find most useful in decision strategies (performance validity) is a completely subject assessment. Design managers were presented 3 representations of the aggregated data – aggregated numerical values of minimum, most likely and maximum, as presented in Appendix C; expert calibrated assessments layered onto one graph as presented in Appendix D; and the BESTFIT[®] aggregated responses represented in Appendix D. The discipline design managers were asked to review each of the representations and comment. One design manager believes all three representations are valuable in that each presents a different interpretable reference to the aggregated data. The numerical representation (Appendix C) provides more of a discrete value to the aggregation while the representation of

layered calibrated assessments allows the design manager to assess the level of agreement between the experts. The BESTFIT[®] aggregated representation is valuable “when you are ready to implement a decision”. Another design manager focused in more on the aggregated graphical representation and responded this representation gave him a better idea of where to go with his decision strategies instead of relying on “heavy interpolation”. In the design manager’s opinion, the aggregated response provides mathematical validity to the “eye-ball” method which is so prevalent in conceptual design. Additionally he felt the methodology took “amalgamous data and transformed it into something useful”.

Each design manager also commented on the usefulness and applicability of this methodology beyond the test case. One of the operations support design managers hopes to extend this methodology to a full scale conceptual vehicle, integrating each discipline necessary to develop a space transport concept. A weights and sizing design manager asserts this methodology is transportable to planetary exploration projects and would like to see this methodology employed in other projects within the Vehicle Analysis Branch at NASA Langley Research Center.

DISCUSSION

As technology systems continue to evolve in complexity and conceptual designers seek to stretch the limits of feasibility of engineering design, strategies to holistically capture and represent risk associated with such highly uncertain and high consequence enterprises becomes paramount. In order to quantify risk with confidence, better quantification strategies need to be developed (Proffitt, 2003). Coupled with improved quantification strategies is the need for more robust combination methods when multiple uncertainties for the same variable have been quantified. Many aggregation methods exist that combine expert opinions when past data is available from which to ascertain likelihood functions and expert credibility. The current research case is embedded in a domain in which these two factors are absent and, therefore, a combination method which does not rely on likelihood functions and with an alternate way to determine expert credibility has been developed.

The development of this methodology for uncertainty assessment required the integration of many elements; a thorough understanding of the research domain and uncertainties under investigation, the development of an efficient, relevant and appropriate data elicitation mechanism, an aggregation approach compatible with the type of uncertainty being investigated and the decision model chosen, and the validation methodology had to be congruent with a subjective research paradigm.

The aerospace conceptual design environment is a unique domain in that there is very limited "hard" data to base decisions on and the preponderance to fully know the outcome of an event is difficult. Aerospace conceptual designs are normally initiated 20-30 years from the time the vehicle needs to be in service and the technology projected to

be incorporated into these new designs is unique and revolutionary. For this very reason, the values of design variables and the uncertainties associated with those values is considered unknown. These values are not completely unknowable since in 20-30 years the vehicles should come to fruition, however at the time of a decision milestone the values are not known quantities. This distinction is important in understanding the research domain and uncertainties under investigation in this study.

The data elicitation mechanism used in the current case was a modification of a questionnaire developed by Monroe (1997). The Monroe questionnaire was chosen as a base model because it had been developed and deployed in the conceptual aerospace environment with success in estimating the Weight Estimating Relationships (WERs) for weights and sizing variables. However, in examining the compatibility of the Monroe questionnaire in the deployment of the aggregation methodology, several issues arose. The structure of the Monroe questionnaire allowed the experts to contradict their own assessments of uncertainty and many of the participants of that study did not find the questionnaire time efficient. None of these complications were too great to overcome but a modification of the questionnaire was necessary to ensure time efficiency and prevent expert contradiction of their own assessments. A strength of the questionnaire for application to the current research is the elicitation of values in triangular distribution form. Triangular distributions work remarkably well when knowledge of the “shoulders” of a distribution is unclear. A weakness of the Monroe structure however, was that he disallowed skewed triangular distributions; the structure of the questions forced experts to give a symmetrical answer around the mean. For this study, experts must be allowed to provide skewed distributions to properly capture the uncertainties of a variable of

interest. Physical limitations of a variable may constrain an extreme value (minimum or maximum) thus impacting the uncertainty associated with a mean value. The questionnaire was modified to allow for skewed distributions to acknowledge possible physical limitations of design variables.

Relating to the uncertainty assessments of the experts is the question of how to assign credibility factors to each expert's assessment in the absence of historical data or observable outcomes. The expert weighting factors are a critical component to the opinion pool aggregation algorithms. Some researchers assign credibility factors based upon years of experience or educational level but as the literature indicates, these characteristics are not necessarily qualifiers for expertise. For this research study, it was determined that eliciting credibility factors from the design managers themselves, would be the most appropriate method to gather these factors. This method of subjectively ascertaining credibility factors would incorporate all the relevant characteristics of expertise and provide the most comprehensive weighting of the experts judgments.

The execution of the linear opinion pool aggregation algorithm provides a "most likely" value of the combined distributions. For a more robust representation of the combination algorithm, simulation sampling was incorporated. The Latin Hypercube sampling technique was run with 500, 1000, 10000, and 15000 iterations; 15000 iterations was the convergence point of the majority of variables. It is interesting to note that on almost all variable sampling, the smaller iterations resulted in either normal or lognormal distributions while the higher iterations (10000 and 15000) converged to a beta distribution. It would be interesting to extend this research to investigate the correlation

between input distributions and output distributions and to hypothesize on the relationship between the two.

Perhaps the biggest challenge in the development of the present research was determining a means to validate the methodology. The distant-future nature of aerospace technology impact rendered classic (objective) validation techniques moot. A means to subjectively validate the usefulness of the aggregated responses to decision-makers was necessary. To assess how useful decision-makers found the aggregated responses, a validation technique that allowed free exchange of ideas, thoughts, criticisms and opinions was necessary. The use of an unstructured interview process provided greater depth of detail in the validation analysis; much more detail than if a questionnaire instrument were used. When eliciting qualitative assessments in a questionnaire, participants may not provide as detailed a response as a researcher might like. Frary (unknown) discourages open-end questions on a questionnaire because participants are likely to suppress responses to save on time commitment to the process. Conversational, context driven dialogue enabled a more comprehensive evaluation of the usability of the aggregated response to conceptual decision-makers.

CONCLUSION

This work began through motivation from expert elicitation applications that resulted in widely disparate evaluations of advanced technology impact on future aerospace vehicle design, performance and operations. Through an extensive review of literature on uncertainty, expert judgment elicitation, calibration, and aggregation, a methodology has been developed that utilizes characteristics of an expert elicitee to adjust the rendered judgments. These adjusted, or calibrated judgments lead to uncertainty distribution that provided a more consistent response among multiple experts in analytical modeling of aerospace vehicle concepts, performance and cost.

The development and deployment of an aggregation methodology has resulted in a process that permits aggregation of multiple expert opinions into a single consensus distribution. While research and application of aggregation techniques is not new, the development of an aggregation methodology in the absence of likelihood functions and expert credibility assessments is unique. The use of simulation in the aggregation process expands aggregation from a single point outcome, generally a most likely value to the aggregation of distributions which more effectively represent risk and uncertainty. The application of a distribution-fitting tool such as BESTFIT® adds a level of robustness to the aggregation outcome by allowing an analyst to vary distribution types to compare aggregated outcome ranges.

This methodology, applied in the aerospace discipline, which is very dynamic and continuously evolving, should prove an effective aid to decision-making associated with aerospace development.

EXTENSIONS OF RESEARCH

The present study has great potential for future expansion. In particular, this research has utilized the triangular distribution for its expert elicitation process. This follows previous work in expert elicitation related to aerospace conceptual design environments (Hampton, 2001; Monroe, 1997). Extending the elicitation process to include other representations of uncertainty assessments such as a beta or exponential distribution may enhance the robustness of the aggregation methodology.

Three additional areas for possible future research in extension or application of the methodology are noted here. First, the addition of experts from the fields of construction and operation of aerospace vehicles as similar as possible to the concepts being studied could bring fresh perspectives to some of the analyses, particularly in the assessments of impacts on performance and the lessening of uncertainty about operational- or performance-related parameters. Second, the potential exists for utilizing the current approach as a means of developing expertise in newer practitioners in a field or discipline. A database containing the questionnaires, calibrations, feedback processes and applications, can be very useful in documenting the results and help compress the learning experience for future experts. Third would be the development of an automated web-based questionnaire administration and database tool. The Web-based survey software, Inquisite® (Catapult Systems, Austin TX), has been used in this study to develop a web-based version of a sample expert judgment elicitation questionnaire and appears to be a good starting point (Task Report -5).

As mentioned in Task Report 6, the logarithmic opinion pool method is also a viable mathematical aggregation method to deploy in subjective probability combination

schemes. Advancing this research to include the logarithmic opinion pool method and evaluating the fidelity of results with the linear opinion pool method would be an interesting and valuable comparative analysis.

Lastly, extension of the methodology developed herein could be applied to domains outside the aerospace conceptual design environment. Many decision domains such as military intelligence, defense systems, national security/terror analysis and many industrial fields such as medical/pharmaceutical fields consistently deal with highly uncertain, high consequence decisions. Extension of this methodology to other decision domains may serve to demonstrate the methodologies use as a template and add to the generalizability of this research.

In addition to the potential extensions and refinement of the methodology developed herein, it is recommended that the methodology be applied to a new space transportation concept being developed. The techniques presented may provide the data necessary to assess uncertainty and risk and search for a robust solution.

Appendix A: Expert Judgment Sample Questionnaire (Operations and Support)

From the RMAT INPUT parameters you have

Variable Scheduled

Nominal 114750

Rate the degree of uncertainty that you associate with this

Low Low/moderate Moderate Moderate/high High

Uncertainty Rating

If you feel this INPUT parameter's default value should be modified, you may provide a new estimate for the INPUT parameter's nominal

New Nominal Value

If you feel the range of possible values around the nominal value is not symmetrical, please provide own estimates of minimum and maximum

Min Max

Now that you have rated the uncertainty for this INPUT parameter, please provide a reason or for your rating. Include a rationale for any change you made to the parameter's nominal

To further document your thinking, please provide any cues (or triggers) that influence your about this

After completing the preceding steps for all parameters you have rated as uncertain, please provide a quantitative explanation of your understanding of Low, Moderate and High uncertainty, using the 5-point scales provided.

The amount of uncertainty or variation that I associate with Low Uncertainty is:

LOW Uncertainty
Less 5% 7.50% 10% 12.50% 15% More

The amount of uncertainty or variation that I associate with Moderate Uncertainty is:

MODERATE Uncertainty
Less 10% 15% 20% 25% 30% More

The amount of uncertainty or variation that I associate with High Uncertainty is:

HIGH Uncertainty
Less 20% 30% 40% 50% 60% More

Appendix B: Calibrated Distributions Operations & Support

TSTO Mach 3 Booster

Variable	Var Name	Nom Value	Expert A Calibrated Distribution			Expert B Calibrated Distribution			Expert C Calibrated Distribution		
			a2	c2	b2	a2	c2	b2	a2	c2	b2
VAR 01	Sched Hrs	114750	40000.00	120487.50	125000.00	90000.00	135460.36	140000.00	90000.00	100000.00	110000.00
VAR 04	MHMA Cal	1	0.25	1.05	1.25	0.25	1.18	1.50	0.25	1.00	1.50
VAR 05	FractsequentWk	0.05	0.02	0.05	0.08	0.03	0.06	0.06	0.05	0.10	0.15
VAR 06	GroundProc	7689	150.00	672.00	1250.00	150.00	755.51	1250.00	150.00	640.00	1250.00
VAR 07	TargMinVehProcDays	30	30.44	31.50	32.56	33.58	35.41	37.25	27.00	30.00	33.00
VAR 08	LPadTimeDays	25.3	15.00	26.57	45.00	15.00	29.87	45.00	15.00	25.30	45.00
VAR 09	No.CrewsAss/shift	1	1.00	1.05	3.00	1.00	1.18	3.00	1.00	1.00	3.00
VAR 10	MTBMCaI	0.833	0.70	0.87	1.40	0.60	0.98	1.20	0.60	0.83	1.20
VAR 12	VehIntTimeDays	5.5	2.00	5.78	6.00	6.16	6.49	6.83	4.95	5.50	6.05
VAR 14	CritFailRate	0.0006052	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VAR 15	FractInheFailures	0.1836	0.17	0.19	0.19	0.17	0.22	0.26	0.13	0.18	0.24

Operations & Support TSTO Mach 3 Orbiter

Variable	Var Name	Nom Value	Expert A Calibrated			Expert B Calibrated			Expert C Calibrated		
			a2	c2	b2	a2	c2	b2	a2	c2	b2
VAR 01	Sched Hrs	159897	50000	167892	170000	140000	188756	195000	140000	159897	195000
VAR 04	MHMA Cal	1	0.2500	1.0500	1.2500	0.2500	1.1805	1.5000	0.2500	1.0000	1.5000
VAR 05	FractsequentWk	0.05	0.0200	0.0525	0.0800	0.0300	0.0590	0.0600	0.0500	0.1000	0.1500
VAR 06	GroundProc	7771	160.0000	680.4000	1250.0000	150.000	764.953	1250.000	150.000	648.000	1250.000
VAR 07	TargMinVehProcDay	30	30.4393	31.5000	32.5607	33.5774	35.4145	37.2516	27.0000	30.0000	33.0000
VAR 08	LPadTimeDay	25.3	15.0000	26.5650	45.0000	15.0000	29.8662	45.0000	15.0000	25.3000	45.0000
VAR 09	No.CrewsAss/shif	1	1.0000	1.0500	3.0000	1.1192	1.1805	1.2417	0.9000	1.0000	1.1000
VAR 10	MTBMCaI	0.804	0.6000	0.8442	1.1000	0.6000	0.9491	1.2000	0.6000	0.8040	1.2000
VAR 11	Orbit Time	252	255.6905	264.6000	273.5095	282.0498	297.4816	312.9134	226.8000	252.0000	277.2000
VAR 12	VehIntTimeDay	5.5	2.0000	5.7750	6.0000	6.1558	6.4927	6.8295	4.9500	5.5000	6.0500
VAR 14	CritFailRate	0.0005745	0.0006	0.0006	0.0006	0.0006	0.0007	0.0007	0.0005	0.0006	0.0007
VAR 15	FractInheFailure	0.1645	0.1500	0.1727	0.1800	0.1539	0.1942	0.2345	0.1151	0.1645	0.2139

Weights & Sizing
TSTO Mach 3 Booster

Variable	Var	Nom	Expert A Calibrated			Expert B Calibrated		
			a2	c2	b2	a2	c2	b2
VAR	tow:	1.337	1.602	1.663	1.723	1.638	1.685	1.732
VAR	pwr:	1.04	0.850	0.950	0.950	1.225	1.260	1.295
VAR	cgrow:	0.15	0.100	0.266	0.350	0.151	0.189	0.226
VAR	d_pf:	4.42	5.560	5.881	6.202	5.414	5.571	5.727
VAR	d_lo	71.1	91.211	94.655	98.099	87.150	89.660	92.181
VAR	wos:	65	55.000	70.000	70.000	72.734	81.927	91.119

Weights & Sizing
TSTO Mach 3 Orbiter

Variable	Var	Nom	Expert A Calibrated			Expert B Calibrated		
			a2	c2	b2	a2	c2	b2
VAR 03	tow	1.3113	1.3208	1.3971	1.4733	1.6064	1.6528	1.6991
VAR 12	delv1	348	429.338	463.033	496.728	426.321	438.625	450.928
VAR 13	delv_sa	100	123.373	133.055	142.738	111.899	126.041	140.183
VAR 14	delv_do	366	451.545	486.983	522.421	428.962	461.312	493.662

Appendix C: Aggregation Results

Operations & Support

TSTO Mach 3 Booster

<u>Variable</u>	<u>Var Name</u>	<u>Nom Value</u>			
VAR 01	Sched Hrs	114750			
Weighting factor	Expert	MIN	MOST LIKELY	MAX	MODE
0.4	Expert A	40,000.0	120,487.5	125,000.0	
0.4	Expert B	90,000.0	135,460.4	140,000.0	
0.2	Expert C	90,000.0	100,000.0	110,000.0	
Aggregated Responses		69,162.0	106,782.0	125,877.0	110004
<u>Variable</u>	<u>Var Name</u>	<u>Nom Value</u>			
VAR 04	MHMA Cal	1.0000			
Weighting factor	Expert	MIN	MOST LIKELY	MAX	MODE
0.4	Expert A	0.2500	1.0500	1.2500	
0.4	Expert B	0.2500	1.1805	1.5000	
0.2	Expert C	0.2500	1.0000	1.5000	
Aggregated Responses		0.2500	0.9141	1.3376	0.94331
<u>Variable</u>	<u>Var Name</u>	<u>Nom Value</u>			
VAR 05	FractsequentWk	0.0500			
Weighting factor	Expert	MIN	MOST LIKELY	MAX	MODE
0.4	Expert A	0.0200	0.0525	0.0800	
0.4	Expert B	0.0300	0.0590	0.0600	
0.2	Expert C	0.0500	0.1000	0.1500	
Aggregated Responses		0.0239	0.0602	0.0924	0.06039
<u>Variable</u>	<u>Var Name</u>	<u>Nom Value</u>			
VAR 06	GroundProc	7689.0000			
Weighting factor	Expert	MIN	MOST LIKELY	MAX	MODE
0.4	Expert A	150.0000	672.0000	1250.0000	
0.4	Expert B	150.0000	755.5088	1250.0000	
0.2	Expert C	150.0000	640.0000	1250.0000	
Aggregated Responses		150.0000	699.6700	1,250.0000	700.05
<u>Variable</u>	<u>Var Name</u>	<u>Nom Value</u>			
VAR 07	TargMinVehProcDays	30.0000			
Weighting factor	Expert	MIN	MOST LIKELY	MAX	MODE
0.4	Expert A	30.4393	31.5000	32.5607	
0.4	Expert B	33.5774	35.4145	37.2516	
0.2	Expert C	27.0000	30.0000	33.0000	
Aggregated Responses		30.6516	32.7658	34.7260	32.7735
<u>Variable</u>	<u>Var Name</u>	<u>Nom Value</u>			
VAR 08	LPadTimeDays	25.3000			
Weighting factor	Expert	MIN	MOST LIKELY	MAX	MODE
0.4	Expert A	15.0000	26.5650	45.0000	
0.4	Expert B	15.0000	29.8662	45.0000	
0.2	Expert C	15.0000	25.3000	45.0000	
Aggregated Responses		15.0000	29.2110	45.0000	29.013

Operations & Support
TSTO Mach 3 Booster - Continued

<u>Variable</u>	<u>Var Name</u>	<u>Nom Value</u>			
VAR 09	No.CrewsAss/shift	1.0000			
Weighting factor	Expert	MIN	MOST LIKELY	MAX	MODE
0.4	Expert A	1.0000	1.0500	3.0000	
0.4	Expert B	1.0000	1.1805	3.0000	
0.2	Expert C	1.0000	1.0000	3.0000	
Aggregated Responses		1.0000	1.1697	3.0000	1.625
<u>Variable</u>	<u>Var Name</u>	<u>Nom Value</u>			
VAR 10	MTBMCaI	0.8330			
Weighting factor	Expert	MIN	MOST LIKELY	MAX	MODE
0.4	Expert A	0.7000	0.8747	1.4000	
0.4	Expert B	0.6000	0.9833	1.2000	
0.2	Expert C	0.6000	0.8330	1.2000	
Aggregated Responses		0.6442	0.9433	1.3216	0.93758
<u>Variable</u>	<u>Var Name</u>	<u>Nom Value</u>			
VAR 12	VehIntTimeDays	5.5000			
Weighting factor	Expert	MIN	MOST LIKELY	MAX	MODE
0.4	Expert A	2.0000	5.7750	6.0000	
0.4	Expert B	6.1558	6.4927	6.8295	
0.2	Expert C	4.9500	5.5000	6.0500	
Aggregated Responses		4.4074	5.5176	6.2581	5.8875
<u>Variable</u>	<u>Var Name</u>	<u>Nom Value</u>			
VAR 14	CritFailRate	0.0006			
Weighting factor	Expert	MIN	MOST LIKELY	MAX	MODE
0.4	Expert A	0.0006	0.0006	0.0007	
0.4	Expert B	0.0007	0.0007	0.0008	
0.2	Expert C	0.0005	0.0006	0.0007	
Aggregated Responses		0.000617	0.000666	0.000755	0.000664
<u>Variable</u>	<u>Var Name</u>	<u>Nom Value</u>			
VAR 15	FractInheFailures	0.1836			
Weighting factor	Expert	MIN	MOST LIKELY	MAX	MODE
0.4	Expert A	0.1700	0.1900	0.1900	
0.4	Expert B	0.1718	0.2167	0.2617	
0.2	Expert C	0.1285	0.1836	0.2387	
Aggregated Responses		0.1606	0.1967	0.2306	0.1969

Operations & Support
TSTO Mach 3 Orbiter

<u>Variable</u>	<u>Var Name</u>	<u>Nom Value</u>			
VAR 01	Sched Hrs	159897			
Weighting factor	Expert	MIN	MOST LIKELY	MAX	Mode
0.4	Expert A	50000.0	167892.0	170000.0	
0.4	Expert B	140000.0	188756.0	195000.0	
0.2	Expert C	140000.0	159897.0	195000.0	
Aggregated Responses		101153.0	154519.0	181044.0	158807.0
<u>Variable</u>	<u>Var Name</u>	<u>Nom Value</u>			
VAR 06	GroundProc	7771.0000			
Weighting factor	Expert	MIN	MOST LIKELY	MAX	Mode
0.4	Expert A	160.0000	680.4000	1250.0000	
0.4	Expert B	150.0000	764.9530	1250.0000	
0.2	Expert C	150.0000	648.0000	1250.0000	
Aggregated Responses		142.3190	703.9100	1250.0000	705.4700
<u>Variable</u>	<u>Var Name</u>	<u>Nom Value</u>			
VAR 09	No.CrewsAss/shift	1.0000			
Weighting factor	Expert	MIN	MOST LIKELY	MAX	Mode
0.4	Expert A	1.0000	1.0500	3.0000	
0.4	Expert B	1.1192	1.1805	1.2417	
0.2	Expert C	0.9000	1.0000	1.1000	
Aggregated Responses		1.0509	1.3476	1.8779	1.1139
<u>Variable</u>	<u>Var Name</u>	<u>Nom Value</u>			
VAR 10	MTBMCaI	0.8040			
Weighting factor	Expert	MIN	MOST LIKELY	MAX	Mode
0.4	Expert A	0.6000	0.8442	1.1000	
0.4	Expert B	0.6000	0.9491	1.2000	
0.2	Expert C	0.6000	0.8040	1.2000	
Aggregated Responses		0.6000	0.8794	0.1120	0.8862

Operations & Support
TSTO Mach 3 Orbiter - Continued

<u>Variable</u>	<u>Var Name</u>	<u>Nom Value</u>			
VAR 11	Orbit Time	252.0000			
Weighting factor	Expert	MIN	MOST LIKELY	MAX	Mode
0.4	Expert A	255.6905	264.6000	273.5095	
0.4	Expert B	282.0498	297.4816	312.9134	
0.2	Expert C	226.8000	252.0000	277.2000	
Aggregated Responses		258.7560	275.2330	291.5300	275.2430
<u>Variable</u>	<u>Var Name</u>	<u>Nom Value</u>			
VAR 14	CritFailRate	0.000575			
Weighting factor	Expert	MIN	MOST LIKELY	MAX	Mode
0.4	Expert A	0.0006	0.0006	0.0006	
0.4	Expert B	0.0006	0.0007	0.0007	
0.2	Expert C	0.0005	0.0006	0.0007	
Aggregated Responses		0.00057	0.00063	0.00066	0.00063
<u>Variable</u>	<u>Var Name</u>	<u>Nom Value</u>			
VAR 15	FractInheFailures	0.1645			
Weighting factor	Expert	MIN	MOST LIKELY	MAX	Mode
0.4	Expert A	0.1500	0.1727	0.1800	
0.4	Expert B	0.1539	0.1942	0.2345	
0.2	Expert C	0.1151	0.1645	0.2139	
Aggregated Responses		0.1416	0.1776	0.2131	0.1776

Weights & Sizing
TSTO Mach 3 Booster

Variable	Var Name	Nom Value			
VAR 02	tow:	1.3372			
Weighting factor	Expert	MIN	MOST LIKELY	MAX	Mode
0.3	Expert A	1.6027	1.6632	1.7237	
0.7	Expert B	1.6382	1.6854	1.7327	
Aggregated Responses		1.6242	1.6787	1.7341	1.6787
Variable	Var Name	Nom Value			
VAR 03	pwr:	1.04			
Weighting factor	Expert	MIN	MOST LIKELY	MAX	Mode
0.3	Expert A	0.8500	0.9500	0.9500	
0.7	Expert B	1.2251	1.2604	1.2958	
Aggregated Responses		1.1095	1.1573	1.1962	1.1583
Variable	Var Name	Nom Value			
VAR 08	cgrow:	0.15			
Weighting factor	Expert	MIN	MOST LIKELY	MAX	Mode
0.3	Expert A	0.1000	0.2661	0.3500	
0.7	Expert B	0.1519	0.1891	0.2262	
Aggregated Responses		0.1183	0.2040	0.2649	0.2061
Variable	Var Name	Nom Value			
VAR 09	d_pf:	4.42			
Weighting factor	Expert	MIN	MOST LIKELY	MAX	Mode
0.3	Expert A	5.5601	5.8811	6.2020	
0.7	Expert B	5.4148	5.5710	5.7273	
Aggregated Responses		5.4179	5.6641	5.9153	5.6637

Weights & Sizing
TSTO Mach 3 Booster – Continued

<u>Variable</u>	<u>Var Name</u>	<u>Nom Value</u>			
VAR 10	d_lox	71.14			
Weighting factor	Expert	MIN	MOST LIKELY	MAX	Mode
0.3	Expert A	91.2117	94.6557	98.0998	
0.7	Expert B	87.1509	89.6600	92.1812	
Aggregated Responses		88.0210	91.1617	94.2747	91.1642
<u>Variable</u>	<u>Var Name</u>	<u>Nom Value</u>			
VAR 11	wos:	65			
Weighting factor	Expert	MIN	MOST LIKELY	MAX	Mode
0.3	Expert A	55.0000	70.0000	70.0000	
0.7	Expert B	72.7347	81.9271	91.1195	
Aggregated Responses		67.6310	76.8500	85.4620	76.9380

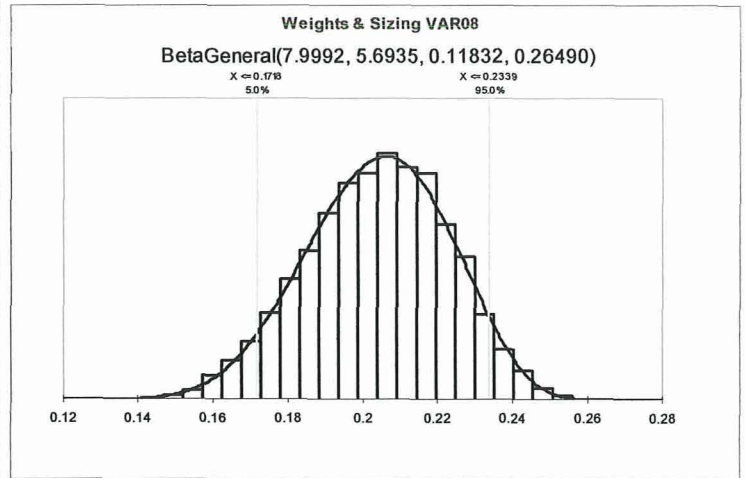
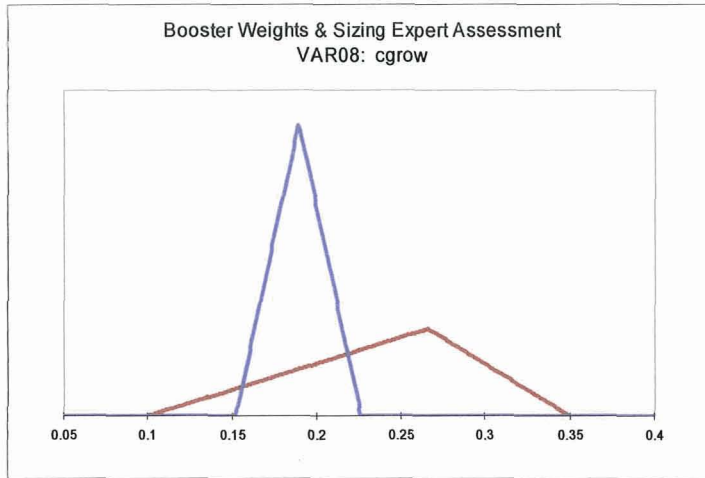
Weights & Sizing
TSTO Mach 3 Orbiter

<u>Variable</u>	<u>Var Name</u>	<u>Nom Value</u>			
VAR 03	tow	1.3113			
Weighting factor	Expert	MIN	MOST LIKELY	MAX	Mode
0.3	Expert A	1.3208	1.3971	1.4733	
0.7	Expert B	1.6064	1.6528	1.6991	
Aggregated Responses		1.5103	1.5761	1.6416	1.57608
<u>Variable</u>	<u>Var Name</u>	<u>Nom Value</u>			
VAR 12	delv1	348			
Weighting factor	Expert	MIN	MOST LIKELY	MAX	Mode
0.3	Expert A	429.3384	463.0333	496.7283	
0.7	Expert B	426.3213	438.6250	450.9287	
Aggregated Responses		426.3313	445.9480	470.4040	445.934
<u>Variable</u>	<u>Var Name</u>	<u>Nom Value</u>			
VAR 13	delv_sa	100			
Weighting factor	Expert	MIN	MOST LIKELY	MAX	Mode
0.3	Expert A	123.3731	133.0556	142.7380	
0.7	Expert B	111.8995	126.0417	140.1838	
Aggregated Responses		115.4150	128.1460	140.7320	128.17
<u>Variable</u>	<u>Var Name</u>	<u>Nom Value</u>			
VAR 14	delv_do	366			
Weighting factor	Expert	MIN	MOST LIKELY	MAX	Mode
0.3	Expert A	451.5455	486.9833	522.4211	
0.7	Expert B	428.9324	461.3125	493.6626	
Aggregated Responses		433.8230	469.0050	505.1810	468.895

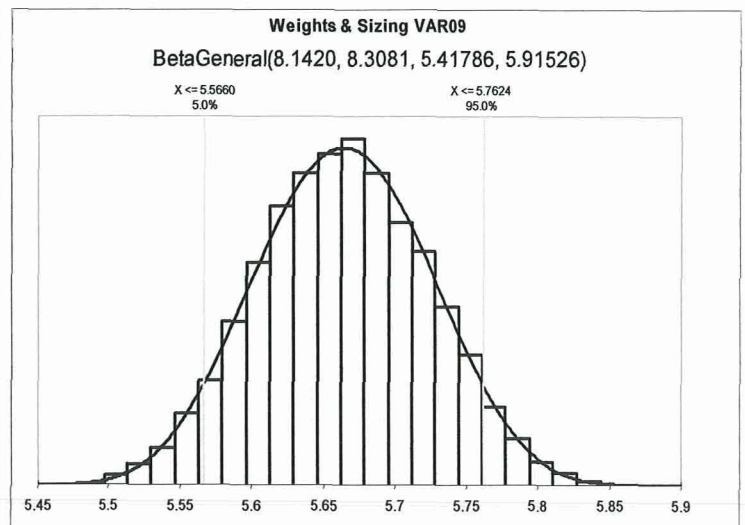
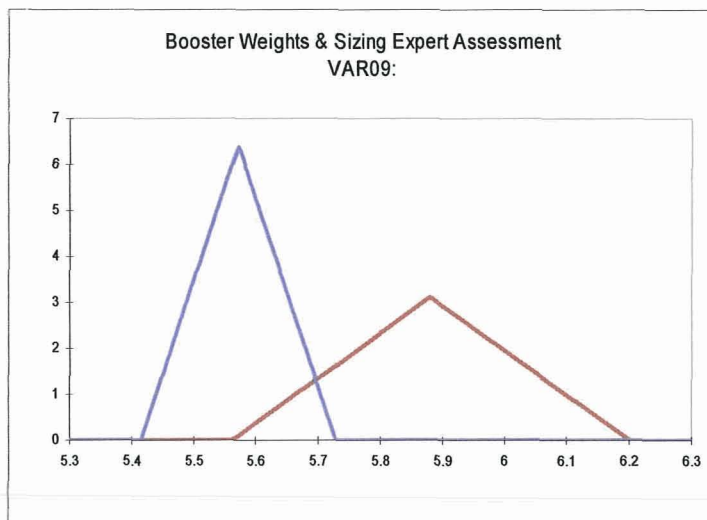
Appendix D: BESTFIT® Distribution by Variable

OPS Booster VAR01	Beta Distribution
OPS Booster VAR04	Beta Distribution
OPS Booster VAR05	Beta Distribution
OPS Booster VAR06	Beta Distribution
OPS Booster VAR07	Beta Distribution
OPS Booster VAR08	Beta Distribution
OPS Booster VAR09	Beta Distribution
OPS Booster VAR10	Beta Distribution
OPS Booster VAR11	Beta Distribution
OPS Booster VAR12	Triangular Distribution
OPS Booster VAR14	Beta Distribution
OPS Booster VAR15	Beta Distribution
OPS Orbiter VAR01	Beta Distribution
OPS Orbiter VAR06	Beta Distribution
OPS Orbiter VAR09	Triangular Distribution
OPS Orbiter VAR10	Weibull Distribution
OPS Orbiter VAR11	Beta Distribution
OPS Orbiter VAR14	Beta Distribution
OPS Orbiter VAR15	Beta Distribution
W&S Booster VAR02	Beta Distribution
W&S Booster VAR03	Beta Distribution
W&S Booster VAR08	Beta Distribution
W&S Booster VAR09	Beta Distribution
W&S Booster VAR10	Beta Distribution
W&S Booster VAR11	Beta Distribution
W&S Orbiter VAR03	Beta Distribution
W&S Orbiter VAR12	Beta Distribution
W&S Orbiter VAR13	Beta Distribution
W&S Orbiter VAR14	Beta Distribution

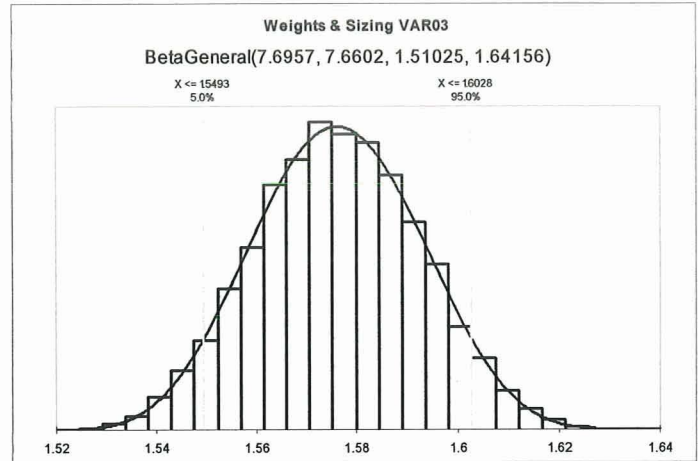
Appendix E: Graphical Representation of Select Aggregated Responses



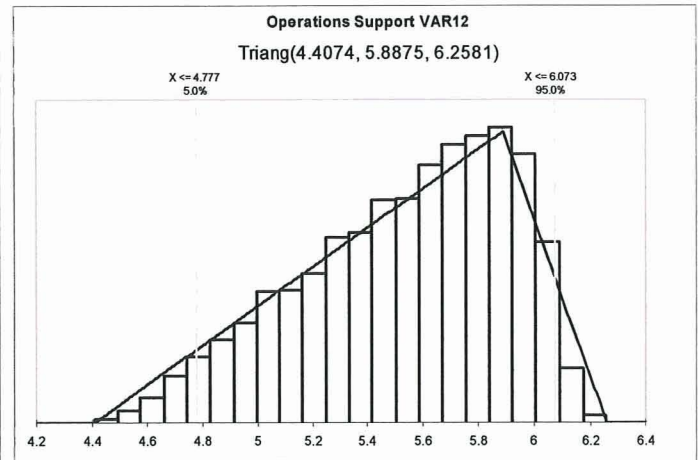
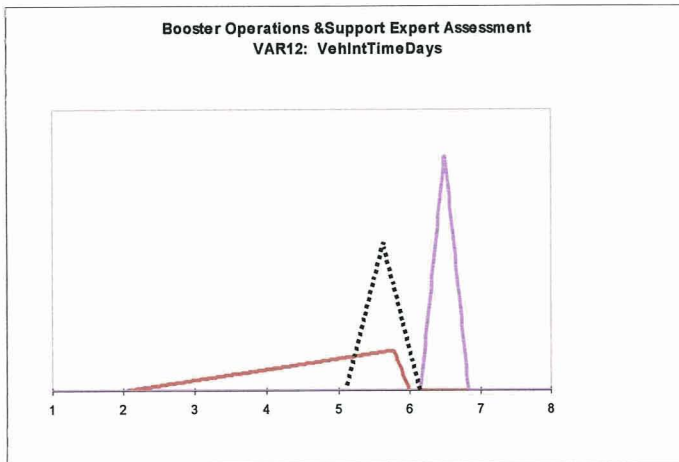
Weights & Sizing Booster VAR08: cgrow



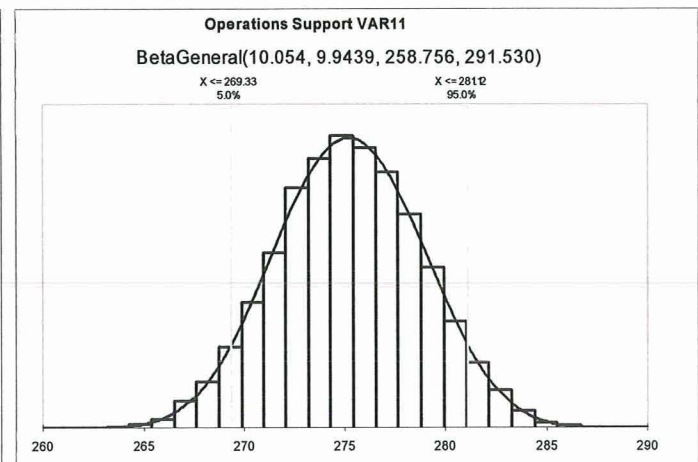
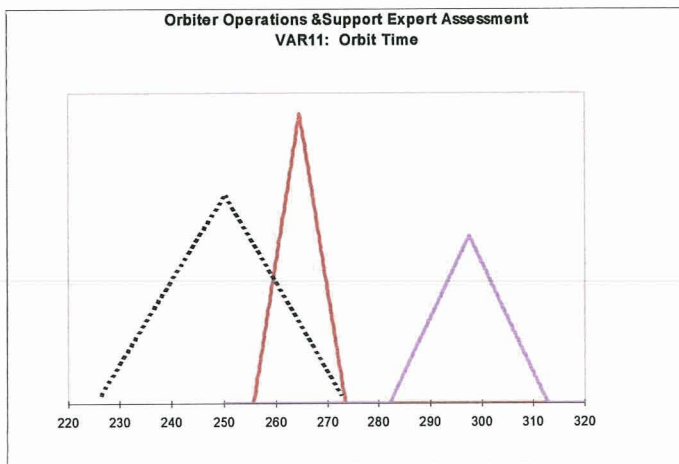
Weights & Sizing Booster VAR09: d_pf



Weights & Sizing Orbiter VAR03: tow



Operations Support Booster VAR12: VehIntTimeDays



Operations Support Booster VAR11: OrbitTime

REFERENCES

- Ashton, Robert H. (1986). Combining the Judgments of Experts: How many and Which Ones? *Organizational Behavior and Human Decision Processes*, Vol. 38, pp. 405-414.
- Ayyub, Bilal, M. (2001). *Elicitation of Expert Opinions for Uncertainty and Risk*. CRC Press, Boca Raton, Florida.
- Barlas, Y. & Carpenter, S. (1990). Philosophical Roots of Model Validation: Two Paradigms. *Systems Dynamics Review*, Vol. 6, No. 2, pp: 148-166.
- Bawcom, D.J. (1997). An Incomplete Handling Methodology for Validation of Bayesian Knowledge Bases. Master's Thesis, Air Force Institute of Technology.
- Braga, R. P. (unknown). Model Validation Methods – Special Reference to Crop Simulation Models. Retrieved October 25, 2003 from http://www.esaelvas.vas.pt/recardo_braga/validation2.pdf
- Clemen, R.T., and Winkler, R.L. (1997). Combining Probability Distributions from Experts in Risk Analysis. *Risk Analysis*, Vol. 19, pp. 187-203.
- Conway, B.A. (2003). *Calibrating Expert Assessments of Advanced Aerospace Technology Adoption Impact*. Ph.D. Dissertation, Old Dominion University, Norfolk, VA. August 2003.
- Crawford, J. D. and L. Stankov (1996). Age Differences in the Realism of Confidence Judgements: A Calibration Study Using Tests of Fluid and Crystallized Intelligence. *Learning and Individual Differences*, Vol. 8, No. 2, 83-103.
- Duarte, B. P. M. 2001. The expected utility theory applied to an individual decision problem -- what technological alternatives to implement to treat industrial solid residuals. *Computers and Operations Research* 28: 357-380.
- Engemann, K.J., Miller, H.E., and Yager, R.R. (1995). *Decision Making Using the Weighted Median Applied to Risk Management*. IEEE Proceedings of The Joint Third International Symposium on Uncertainty Modeling and Analysis-North American Fuzzy Information Processing Society, September 17-20, College Park, Maryland.
- Frery, R.B. (unknown). A Brief Guide to Questionnaire Development. Office of Measurement and Research Service, Virginia Polytechnic Institute and State University.
- Genest, C. and Zidek, J.V. (1986). Combining Probability Distributions: A Critique and an Annotated Bibliography. *Statistical Science*, Vol.1, pp. 114-135
- Haimes, Y.Y (1998). *Risk Modeling, Assessment and Management*. John Wiley & Sons: New York.

- Hampton, K.R. (2001). An Integrated Risk Analysis Methodology in a Multidisciplinary Design Environment. Ph.D. thesis, Old Dominion University
- Hogarth, Robin M. (1978). A Note on Aggregating Opinions. *Organizational Behavior and Human Performance*, Vol. 21, pp. 40-46.
- Hurley, W.J. and Lior, D.U. (2002). Combining Expert Judgment: On the performance of trimmed mean vote aggregation procedures in the presence of strategic voting. *European Journal of Operational Research*, Vol. 14, pp. 142-147.
- Jackson, P. (1999). *Introduction to Expert Systems*. 3rd Edition. Addison-Wesley, Harlow, England.
- Keren, G. (1991). Calibration and probability judgments: Conceptual and methodological issues. *Acta Psychologica*, 77: 217-273.
- Kuhn, T. (1970). *The Structure of Scientific Revolutions*. Chicago, University of Chicago Press.
- Law, A. and Kelton, W. (1991). *Simulation Modeling & Analysis*. 2nd Edition, McGraw-Hill, Inc. New York.
- MacCrimmon, K. R. and D. A. Wehrung (1986). *Taking Risks: The Management of Uncertainty*. New York. The Free Press.
- Mendenhall, W. and Sincich, T. (1995). *Statistics for Engineering and the Sciences*. 4th Edition. Prentice Hall: Upper Sadler River.
- Monroe, R.W. (1997). *A Synthesized Methodology for Eliciting Expert Judgment for Addressing Uncertainty in Decision Analysis*. Ph.D. Dissertation. Old Dominion University, Norfolk, VA. August 1997.
- Morris, W.D. (2003). Personal Communication, NASA Langley Research Center, Hampton, VA., June 24.
- Morgan, M.G., and Henrion, M. (1990). *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*. Cambridge University Press, Cambridge.
- Morris, Peter (1977). Combining Expert Judgments: A Bayesian Approach. *Management Science*, Vol. 23, No. 7, pp. 679-693.
- Morris, Peter (1986). Observations on Expert Aggregation. *Management Science*, Vol. 32, No. 3, pp. 321-328.
- Murphy, A. H. and R. L. Winkler (1974). Credible Interval Temperature Forecasting: Some Experimental Results. *Monthly Weather Review*, 102: 784-794

- Palisade Corporation (2002). *Guide to Using @RISK: Advanced Risk Analysis For Spreadsheets*, Palisade Corporation, Newfield, NY.
- Pederson, K. et al. (2000). The 'Validation Square'-Validating Design Methods. *ASME Design Theory and Methodology Conference*, September, Baltimore, MD, 2000.
- Pennock, David M. (1997). *Market-Based Belief Aggregation and Group Decision Making*. Dissertation Proposal. University of Michigan, Ann Arbor, MI.
- Peterson, R.A. (2000). *Constructing Effective Questionnaires*, Sage Publications, Inc. Thousand Oaks, CA
- Proffitt, T. (2003). Keynote Speaker. *The International Society of Parametric Analysts and The Society of Cost Estimating and Analysis 4th Joint International Conference and Educational Workshop*. June, Orlando, Florida, 2003.
- Rantilla, A.H and Budescu, D.V. (1999). *Aggregation of Expert Opinion*. IEEE Proceedings of the 32nd Hawaii International Conference on System Sciences.
- Rowe, G (1992) Perspectives on Expertise in the Aggregation of Judgments. IN *Expertise and Decision Support*. Wright,, G. and Bolger, Fl editors. Plenum Publishing: New York.
- Rowell, L. F., Braun, R. D., J. R. Olds, and R. Unal (1999). Multidisciplinary Conceptual Design Optimization of Space Transportation Systems. *AIAA Journal of Aircraft*, Vol. 26, No. 1, pp 218-226.
- Sargent, R.G. (1999). Validation and Verification of Simulation Models. *Proceedings of the 1999 Winter Simulation Conference*. December, Phoenix, AZ, 1999.
- Schaefer, R.E. (1976). The Evaluation of Individual and Aggregated Subjective Probability Distributions. *Organizational Behavior and Human Performance*, 17: 199-210.
- Shanteau, J., Weiss, D., Thomas, R. & Pounds, J. (2002). Performance-based assessment of expertise: How to decide if someone is an expert or not. *European Journal of Operational Research*, 136, 253-263.
- Unal, R. and Yeniay, O. (2003). Reducing Design Risk Using Robust Design Methods: A Dual Response Surface. NASA Grant NAG-1-01086 (Old Dominion University Project No: 113091). March 2003.
- Unal, R. (2002). Development Of Response Surface Models And Technology Support Levels. Final Report, NASA PO # H-30938D (Old Dominion University Research Foundation Project No: 11720). February 2002.
- Vose, D. (2000). *Quantitative Risk Analysis*. John Wiley and Sons: Chichester

- Winkler, Robert L. (1981). Combining Probability Distributions from Dependent Information Sources. *Management Science*, Vol. 27, No. 4, pp. 479-488.
- Winkler, Robert L., and Clemen, Robert T. (1992). Sensitivity of Weights in Combining Forecasts. *Operations Research*, Vol. 40, No. 3, pp. 609-614.
- Wright, G., G. Rowe, F. Bolger, and J. Gammack (1994). Coherence, Calibration, and Expertise in Judgmental Probability Forecasting. *Organizational Behavior and human Decision Processes*, 57: 1-25.
- Yager, Ronald R. (1996). *On Mean Type Aggregation*. IEEE Transactions on Systems, Man and Cybernetics, Vol. 26, No. 2, pp. 209-221.